EDA - Exploratory Data Analysis

importing data

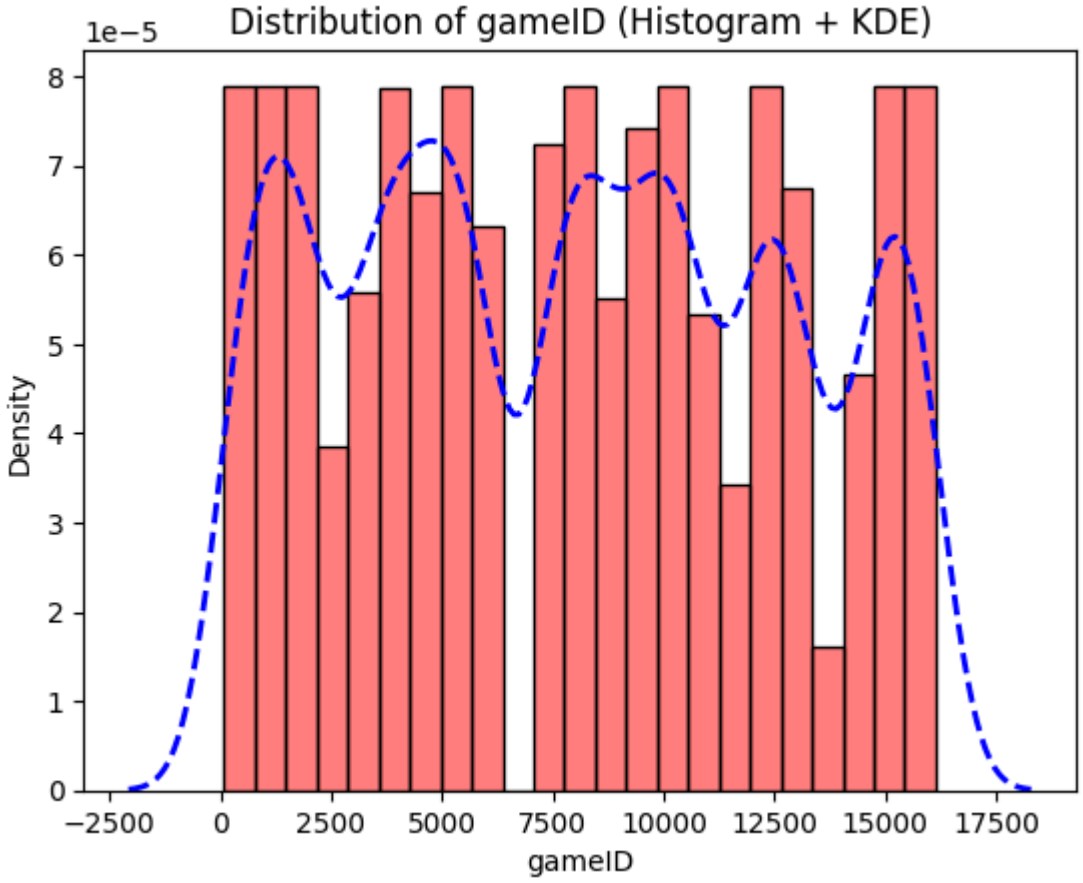
```
Loaded df_appearances from ../pickles/df_appearances.pkl
Skipping player_performance...
Skipping player_game_team_mapping...
Skipping df_games_odds...
Loaded df_teamstats from ../pickles/df_teamstats.pkl
Loaded df_shots from ../pickles/df_shots.pkl
Loaded gameresult from ../pickles/gameresult.pkl
Loaded df_after_outliers_missing from ../pickles/df_after_outliers_missing.pkl
Skipping team_performance...
Loaded df_with_categories from ../pickles/df_with_categories.pkl
Loaded df_num_after_EDA from ../pickles/df_num_after_EDA.pkl
Skipping df_games...
Loaded manipulated_data_no_outleirs from ../pickles/manipulated_data_no_outleirs.p
kl
Loaded player_shots from ../pickles/player_shots.pkl
Loaded df_after_EDA from ../pickles/df_after_EDA.pkl
Loaded teamstats from ../pickles/teamstats.pkl
Loaded df_combined from ../pickles/df_combined.pkl
```

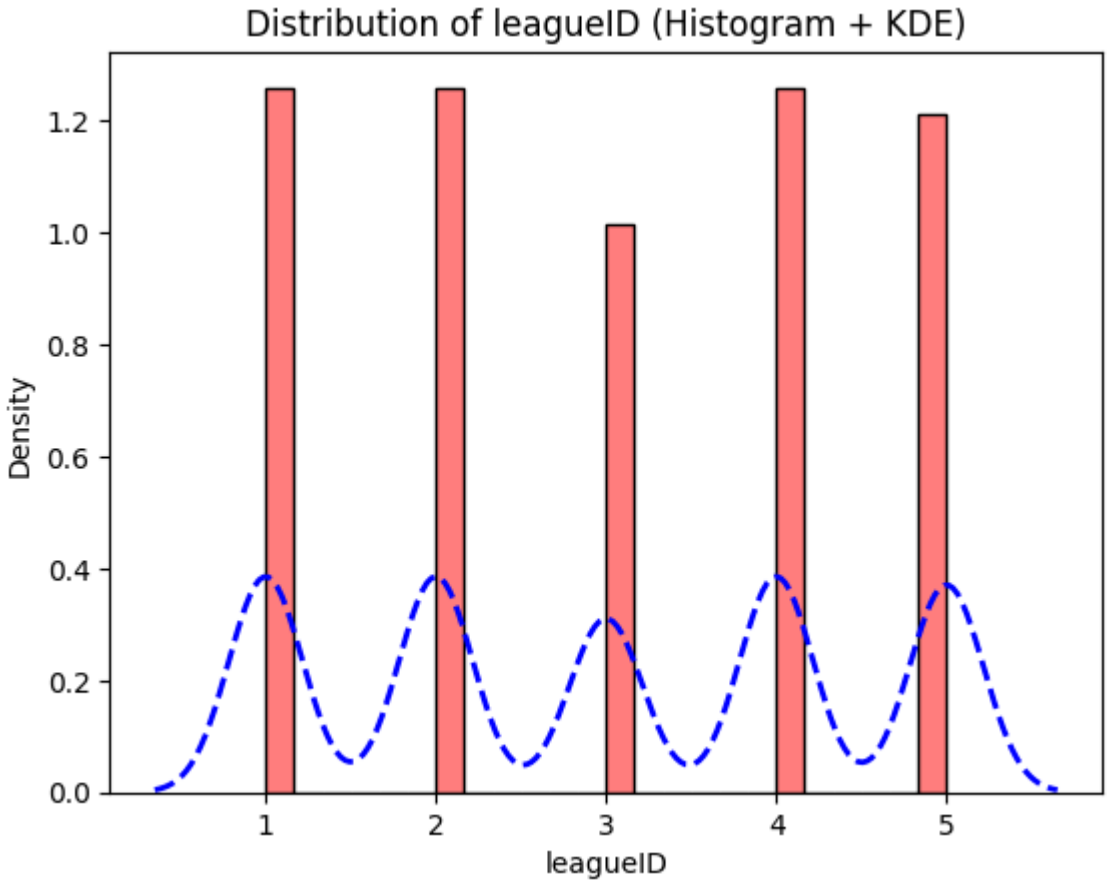
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12680 entries, 0 to 12679
Data columns (total 47 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   gameID                    12680 non-null  int64  
 1   leagueID                  12680 non-null  int64  
 2   season                    12680 non-null  int64  
 3   date                      12680 non-null  object 
 4   homeTeamID                12680 non-null  int64  
 5   awayTeamID                12680 non-null  int64  
 6   homeGoals                 12680 non-null  int64  
 7   awayGoals                 12680 non-null  int64  
 8   homeGoalsHalfTime         12680 non-null  int64  
 9   awayGoalsHalfTime         12680 non-null  int64  
 10  home_xGoals               12680 non-null  float64
 11  home_shots                12680 non-null  int64  
 12  home_shotsOnTarget        12680 non-null  int64  
 13  home_deep                 12680 non-null  int64  
 14  home_ppda                 12680 non-null  float64
 15  home_fouls                12680 non-null  int64  
 16  home_corners              12680 non-null  int64  
 17  home_yellowCards          12679 non-null  float64
 18  home_redCards             12680 non-null  int64  
 19  home_total_assists        12680 non-null  int64  
 20  home_total_xAssists       12680 non-null  float64
 21  home_total_key_passes     12680 non-null  int64  
 22  home_total_xGoalsChain    12680 non-null  float64
 23  home_total_xGoalsBuildup  12680 non-null  float64
 24  home_total_yellow_cards   12680 non-null  int64  
 25  home_total_red_cards      12680 non-null  int64  
 26  home_total_blocked_shots  12677 non-null  float64
 27  home_total_saved_shots    12677 non-null  float64
 28  away_xGoals               12680 non-null  float64
 29  away_shots                12680 non-null  int64  
 30  away_shotsOnTarget        12680 non-null  int64  
 31  away_deep                 12680 non-null  int64  
 32  away_ppda                 12680 non-null  float64
 33  away_fouls                12680 non-null  int64  
 34  away_corners              12680 non-null  int64  
 35  away_yellowCards          12680 non-null  float64
 36  away_redCards             12680 non-null  int64  
 37  away_total_assists        12680 non-null  int64  
 38  away_total_xAssists       12680 non-null  float64
 39  away_total_key_passes     12680 non-null  int64  
 40  away_total_xGoalsChain    12680 non-null  float64
 41  away_total_xGoalsBuildup  12680 non-null  float64
 42  away_total_yellow_cards   12680 non-null  int64  
 43  away_total_red_cards      12680 non-null  int64  
 44  away_total_blocked_shots  12672 non-null  float64
 45  away_total_saved_shots    12672 non-null  float64
 46  gameresult                12680 non-null  object 
dtypes: float64(16), int64(29), object(2)
memory usage: 4.6+ MB
```

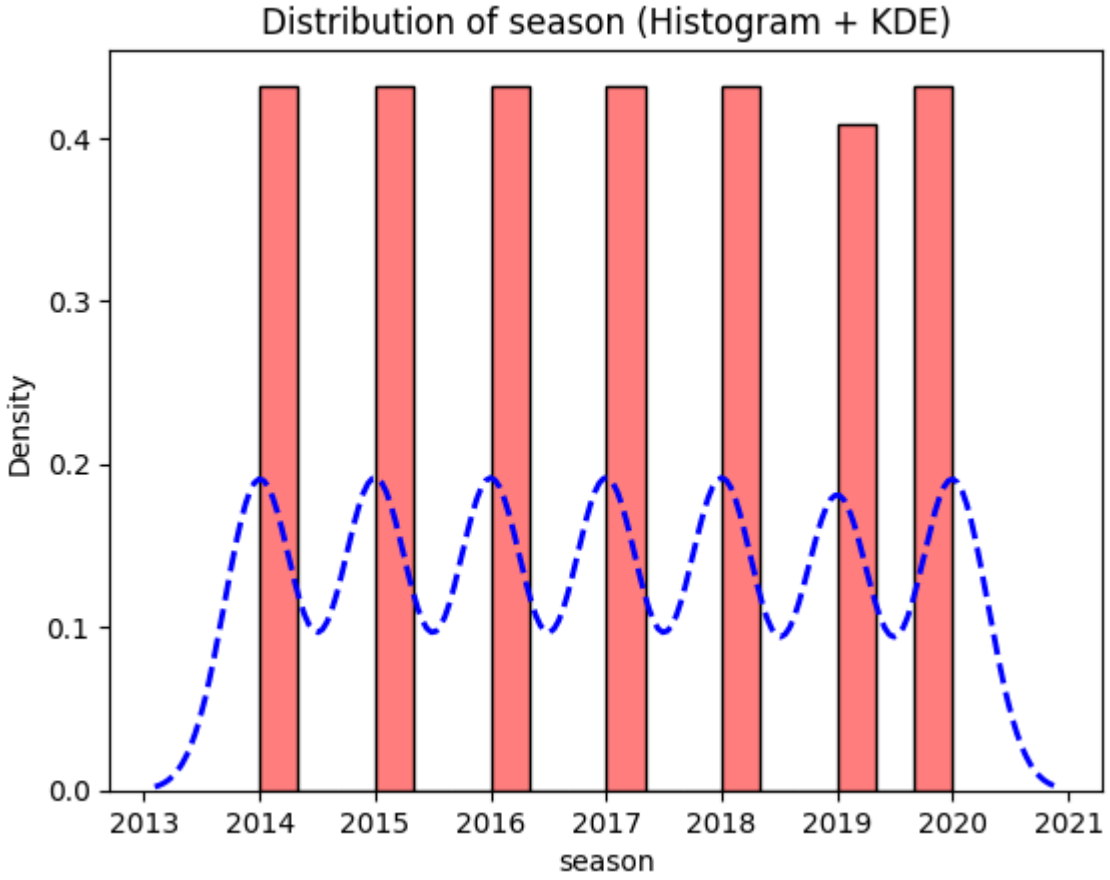
	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
0	81	1	2015	2015-08-08 15:45:00	89	82	1	0
1	82	1	2015	2015-08-08 18:00:00	73	71	0	1
2	83	1	2015	2015-08-08 18:00:00	72	90	2	2
3	84	1	2015	2015-08-08 18:00:00	75	77	4	2
4	85	1	2015	2015-08-08 18:00:00	79	78	1	3
...	...	...	...	...	...	...	...	...
12675	16131	5	2020	2021-05-23 19:00:00	168	166	1	2
12676	16132	5	2020	2021-05-23 19:00:00	177	176	1	2
12677	16133	5	2020	2021-05-23 19:00:00	163	235	2	0
12678	16134	5	2020	2021-05-23 19:00:00	175	181	0	1
12679	16135	5	2020	2021-05-23 19:00:00	225	179	1	1

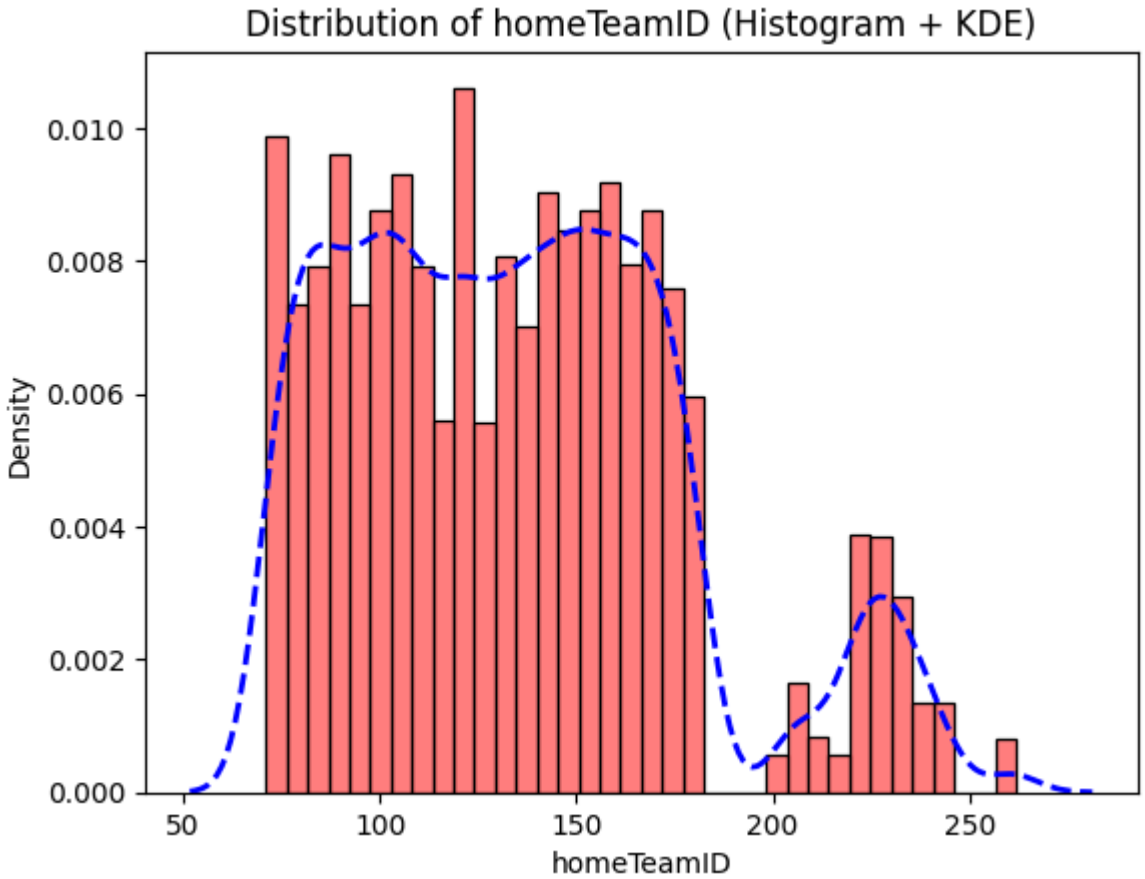
12680 rows × 47 columns

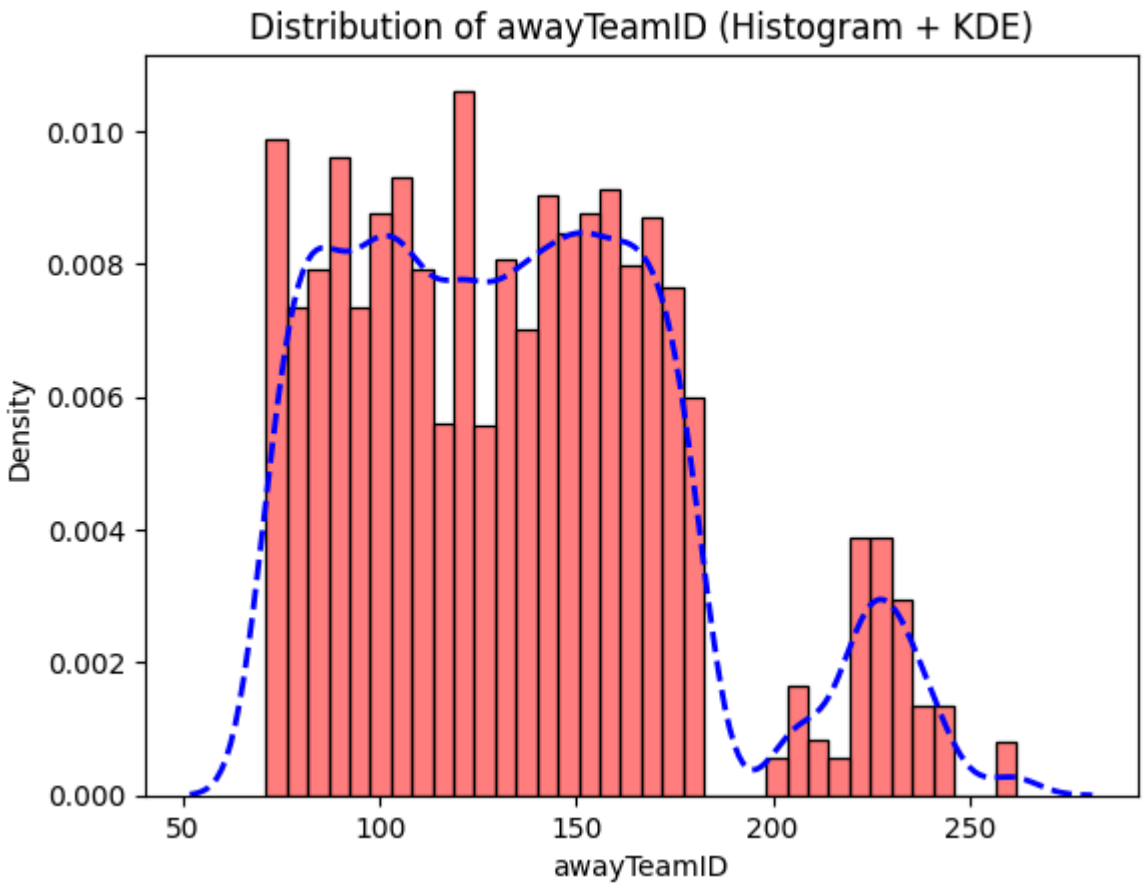
Data visualization

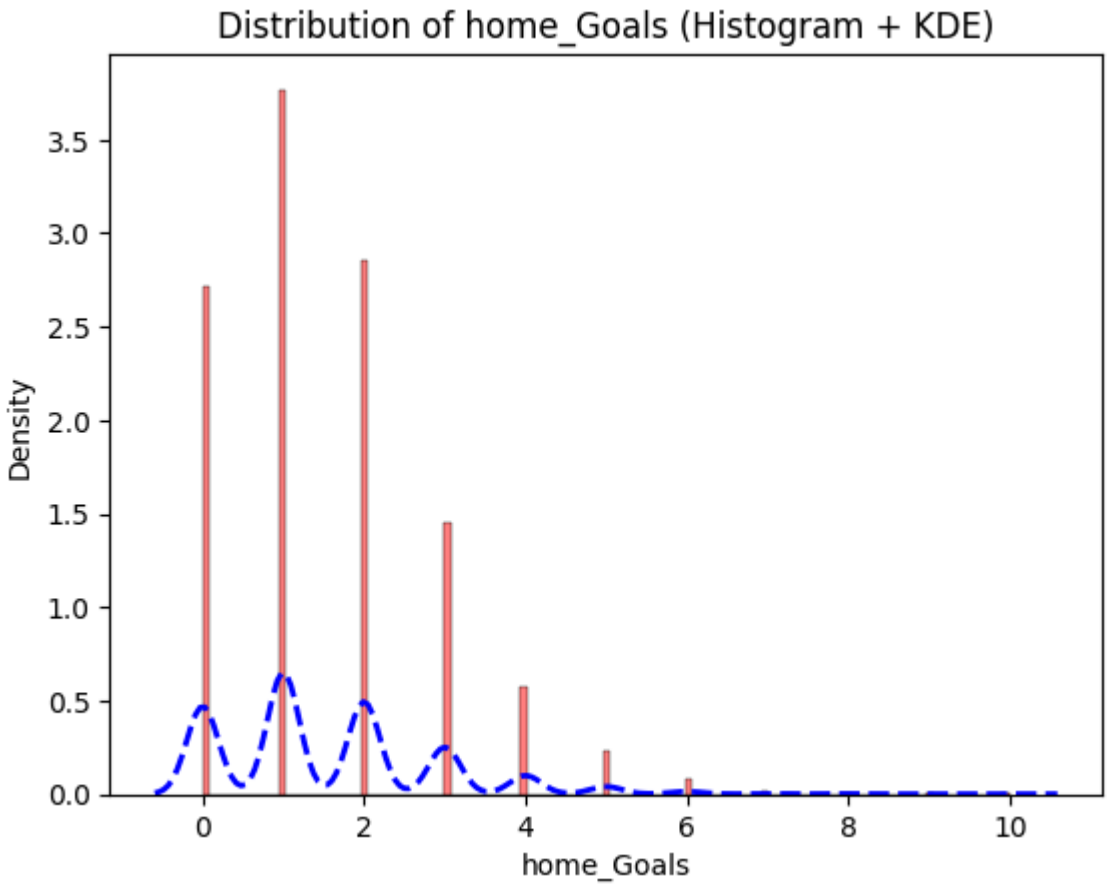

Distribution of gameID (Histogram + KDE)
1e−5
Density
gameID
−2500
0
2500
5000
7500
10000
12500
15000
17500

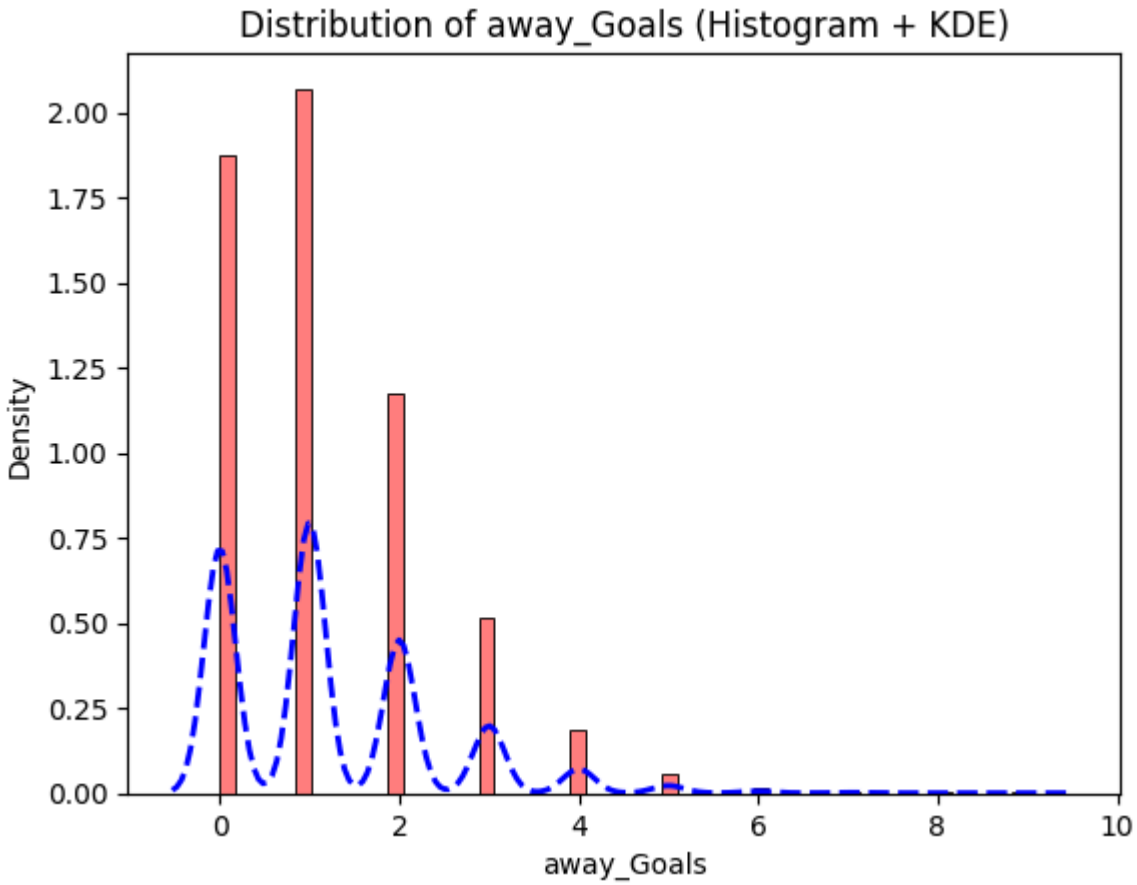

Distribution of leagueID (Histogram + KDE)
Density
leagueID
1
2
3
4
5

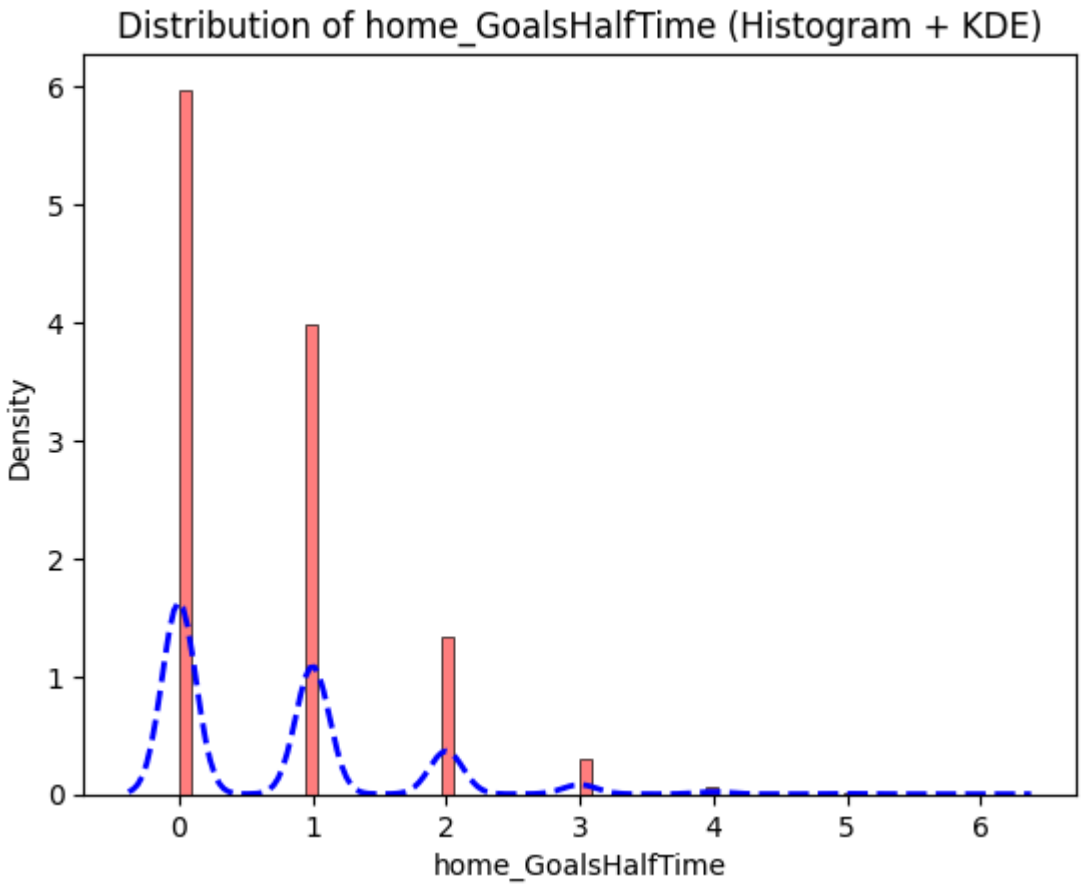

Distribution of season (Histogram + KDE)
Density
0.0
0.1
0.2
0.3
0.4
2013
2014
2015
2016
2017
2018
2019
2020
2021
season

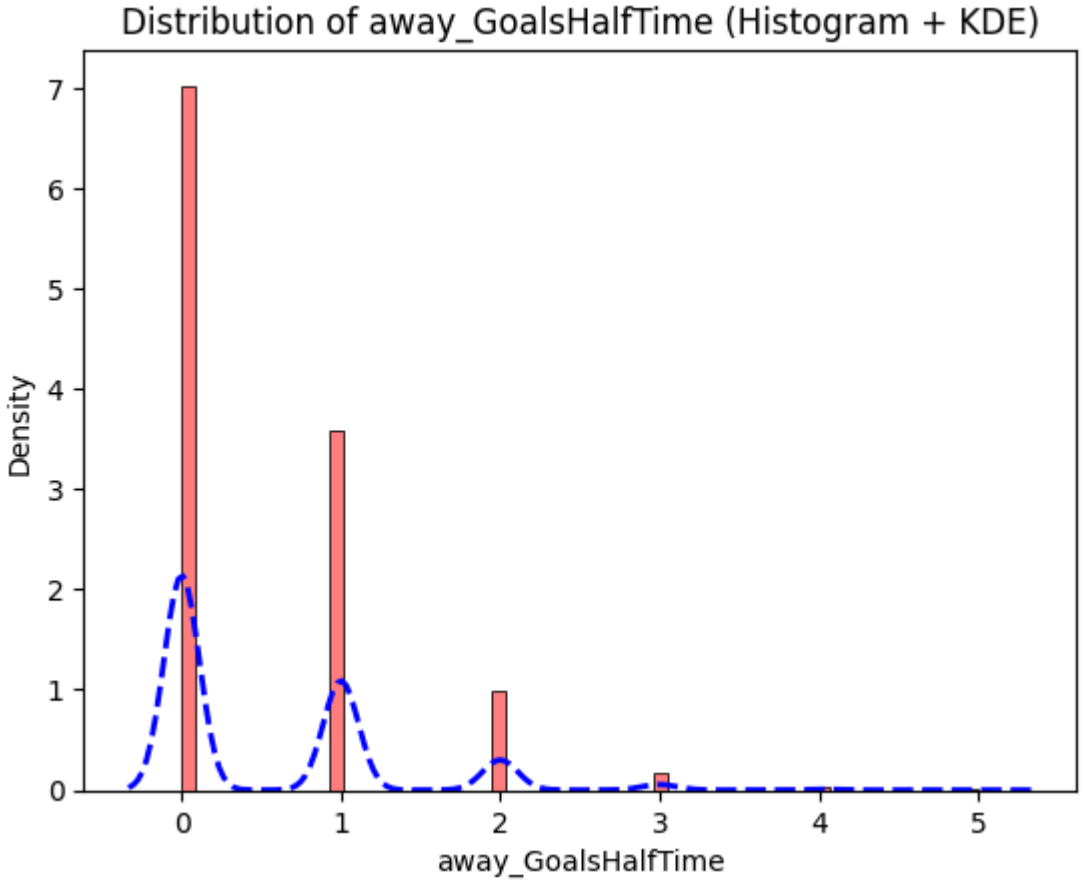

Distribution of homeTeamID (Histogram + KDE)
Density
0.000
0.002
0.004
0.006
0.008
0.010
50
100
150
200
250
homeTeamID

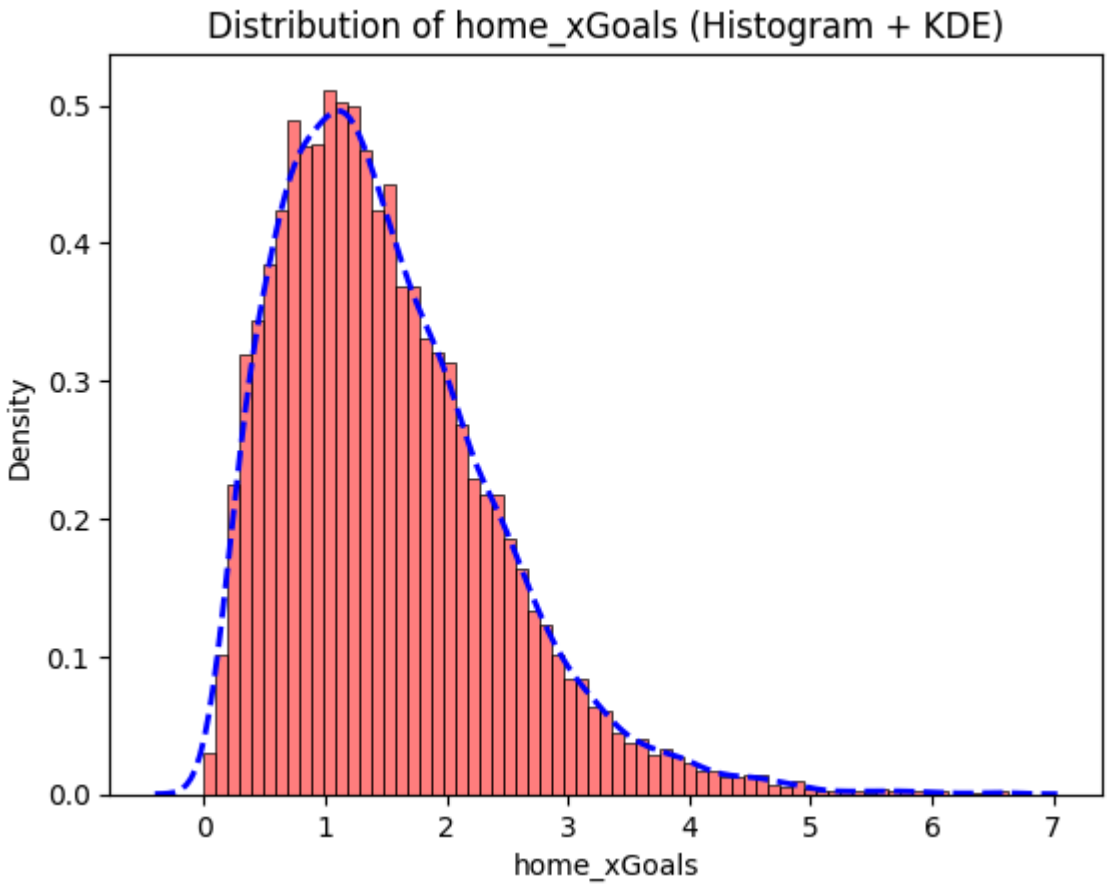

Distribution of awayTeamID (Histogram + KDE)
Density
0.010
0.008
0.006
0.004
0.002
0.000
50
100
150
200
250
awayTeamID

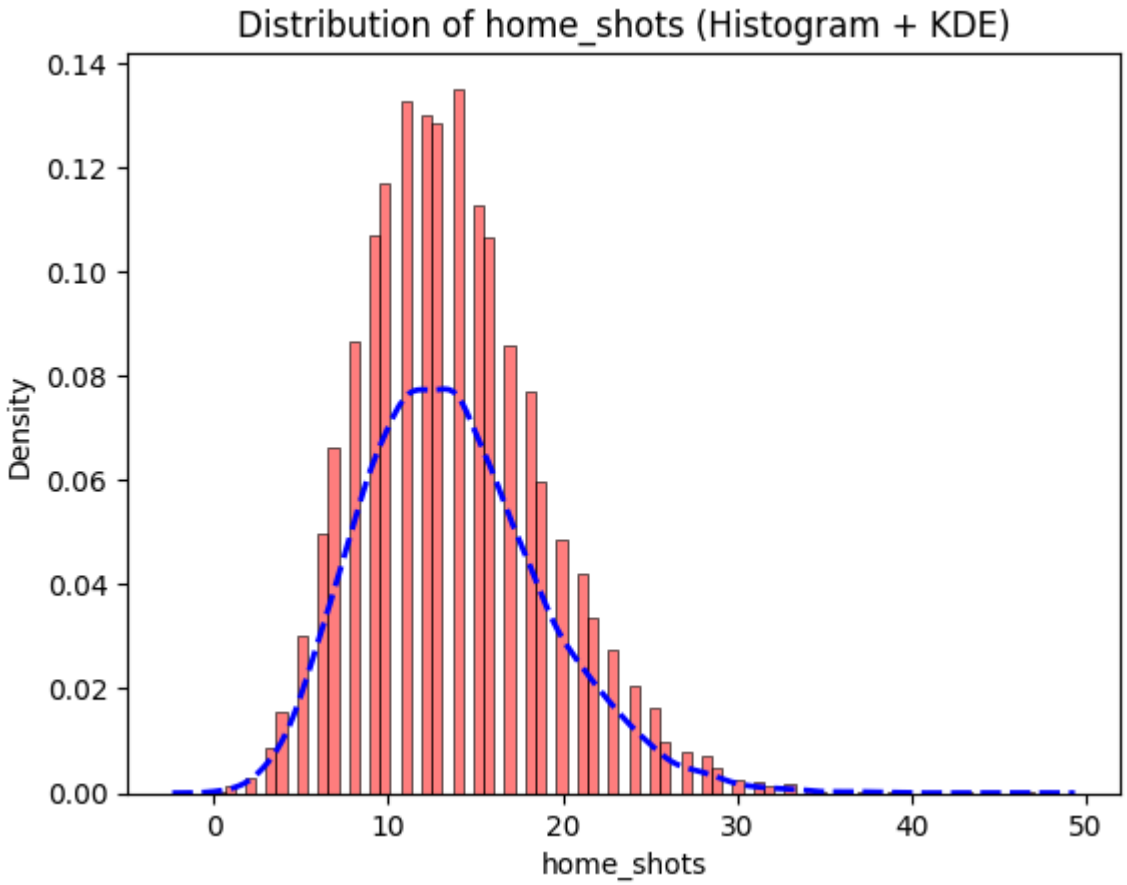

Distribution of home_Goals (Histogram + KDE)
Density
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
10
home_Goals

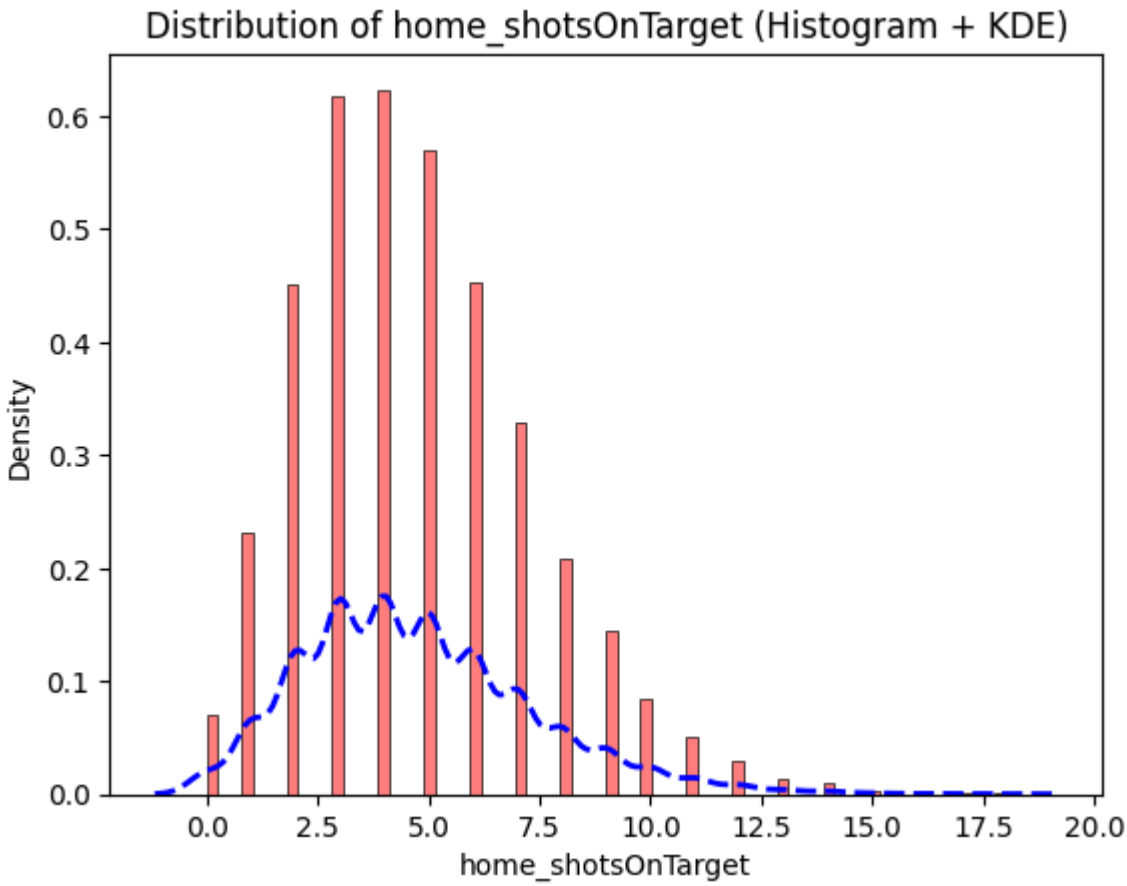

Distribution of away_Goals (Histogram + KDE)
Density
away_Goals
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0
2
4
6
8
10

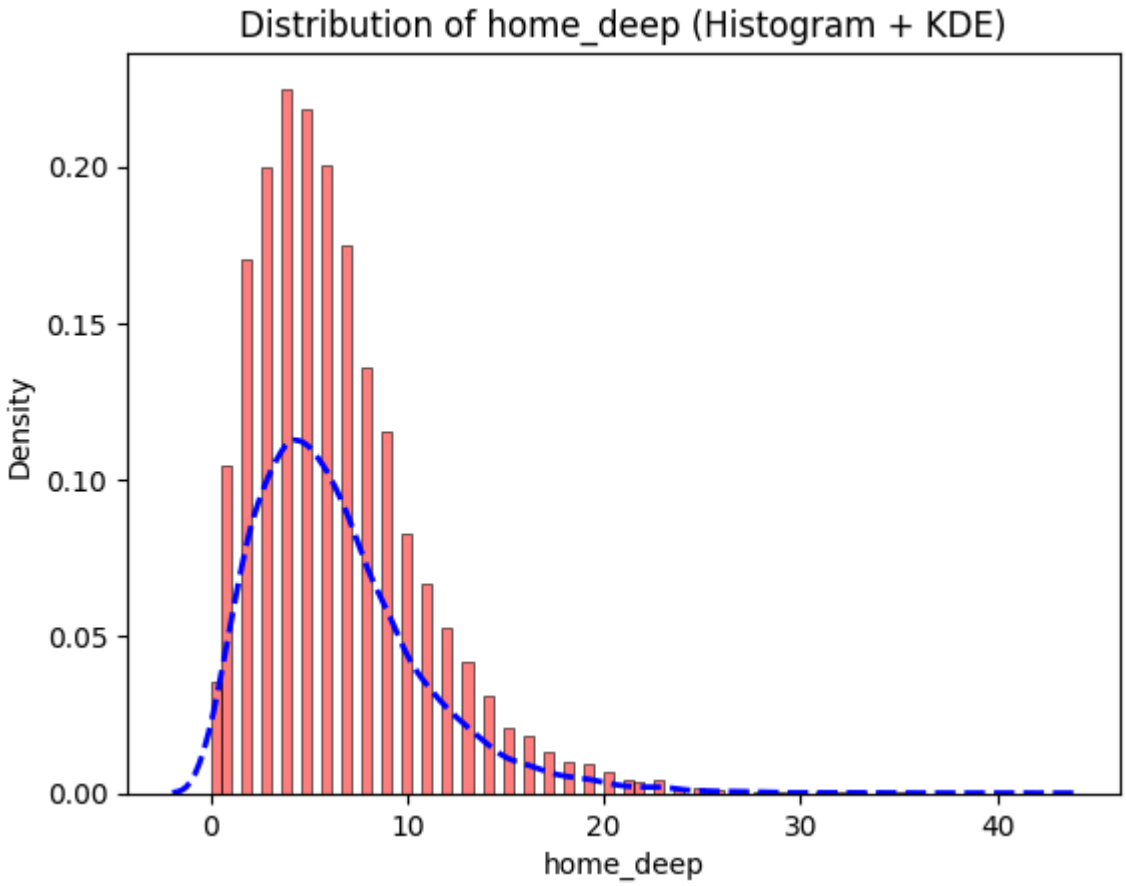

Distribution of home_GoalsHalfTime (Histogram + KDE)
Density
home_GoalsHalfTime
0
1
2
3
4
5
6

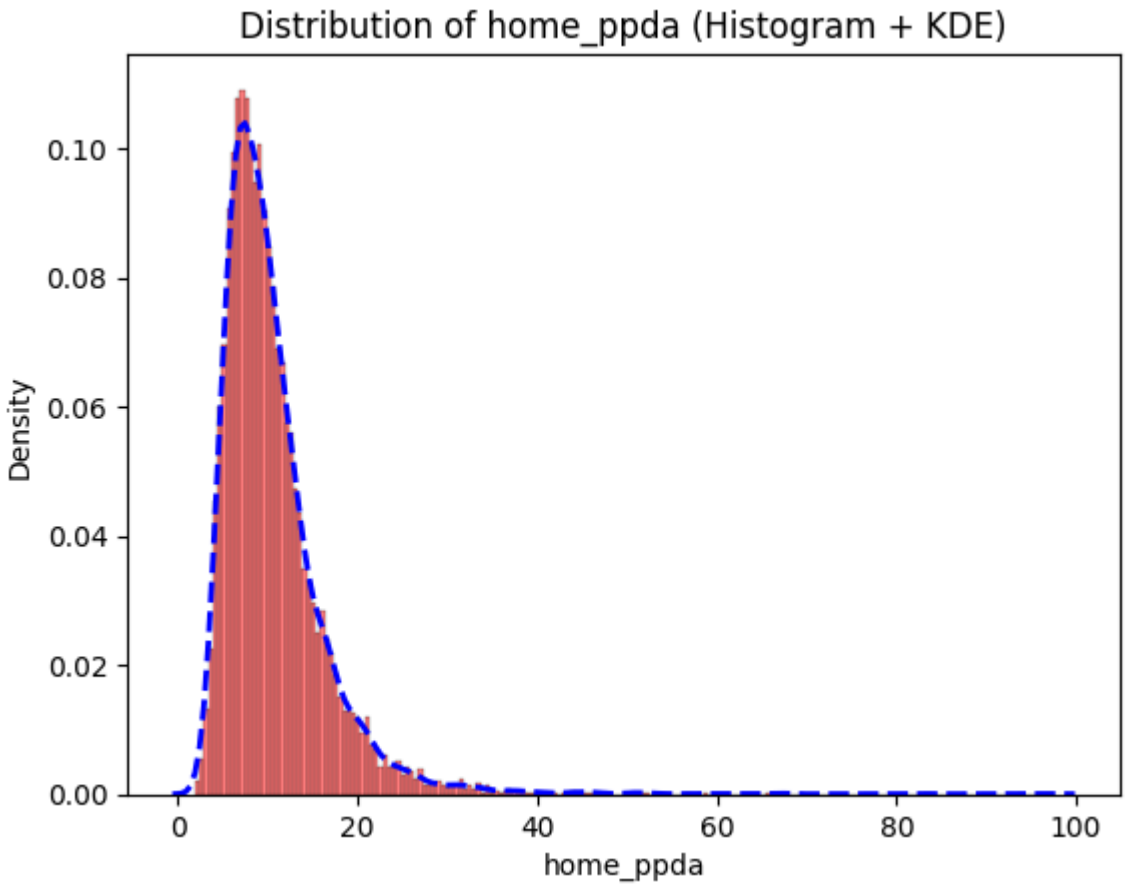

Distribution of away_GoalsHalfTime (Histogram + KDE)
Density
away_GoalsHalfTime
0
1
2
3
4
5
6
7

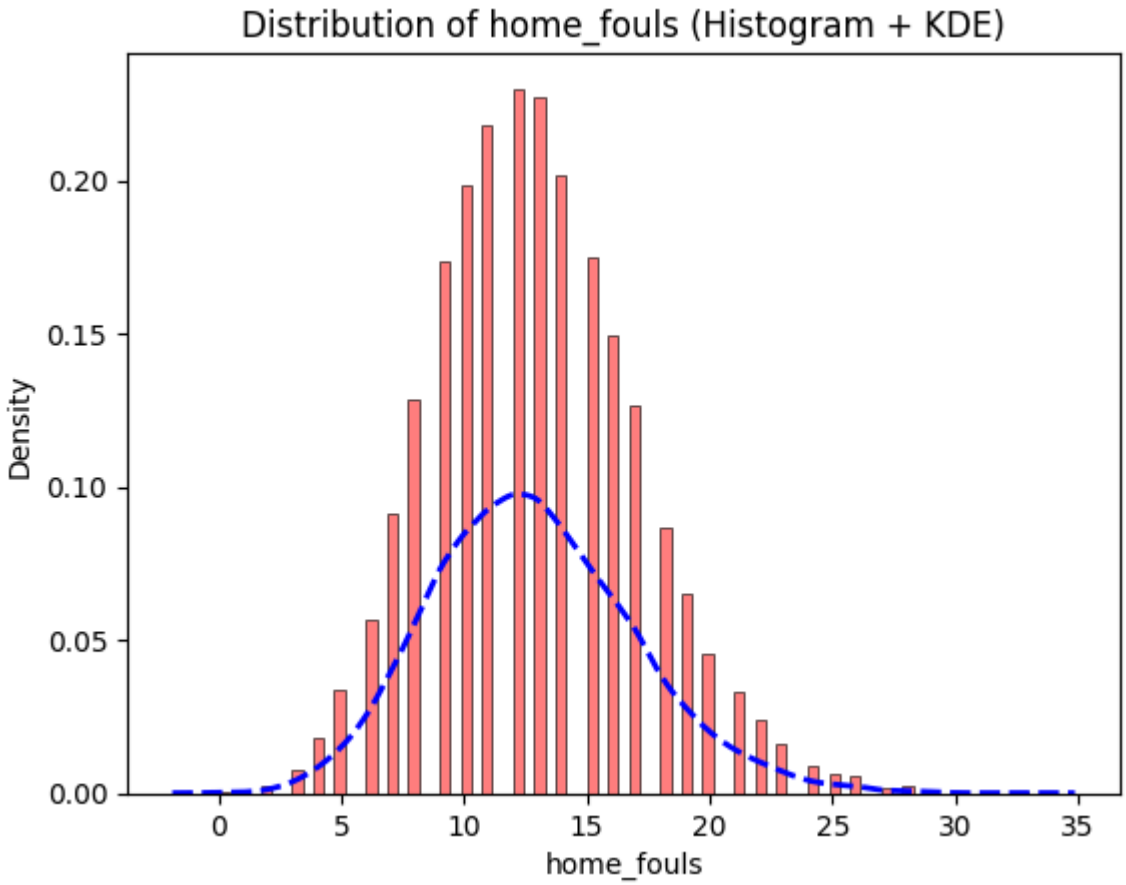

Distribution of home_xGoals (Histogram + KDE)
Density
home_xGoals
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6
7

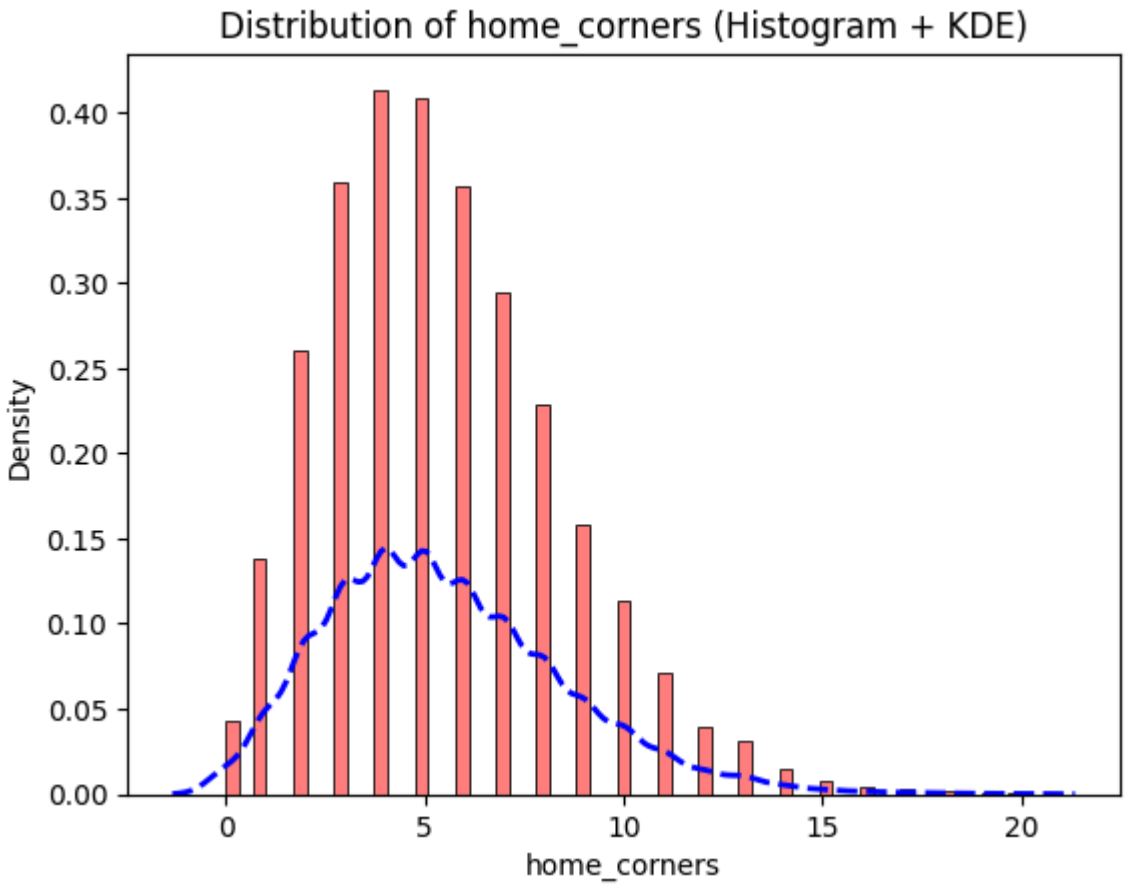

Distribution of home_shots (Histogram + KDE)
Density
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50
home_shots

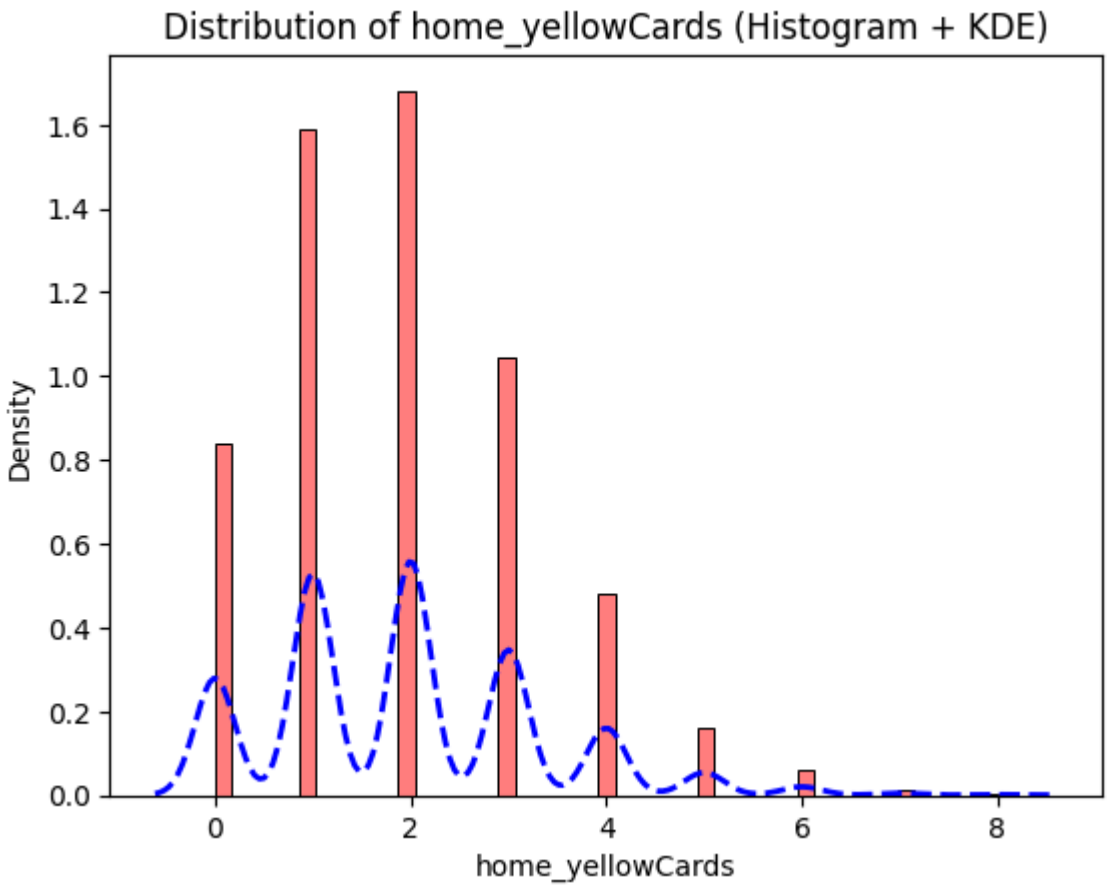

Distribution of home_shotsOnTarget (Histogram + KDE)
Density
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
home_shotsOnTarget

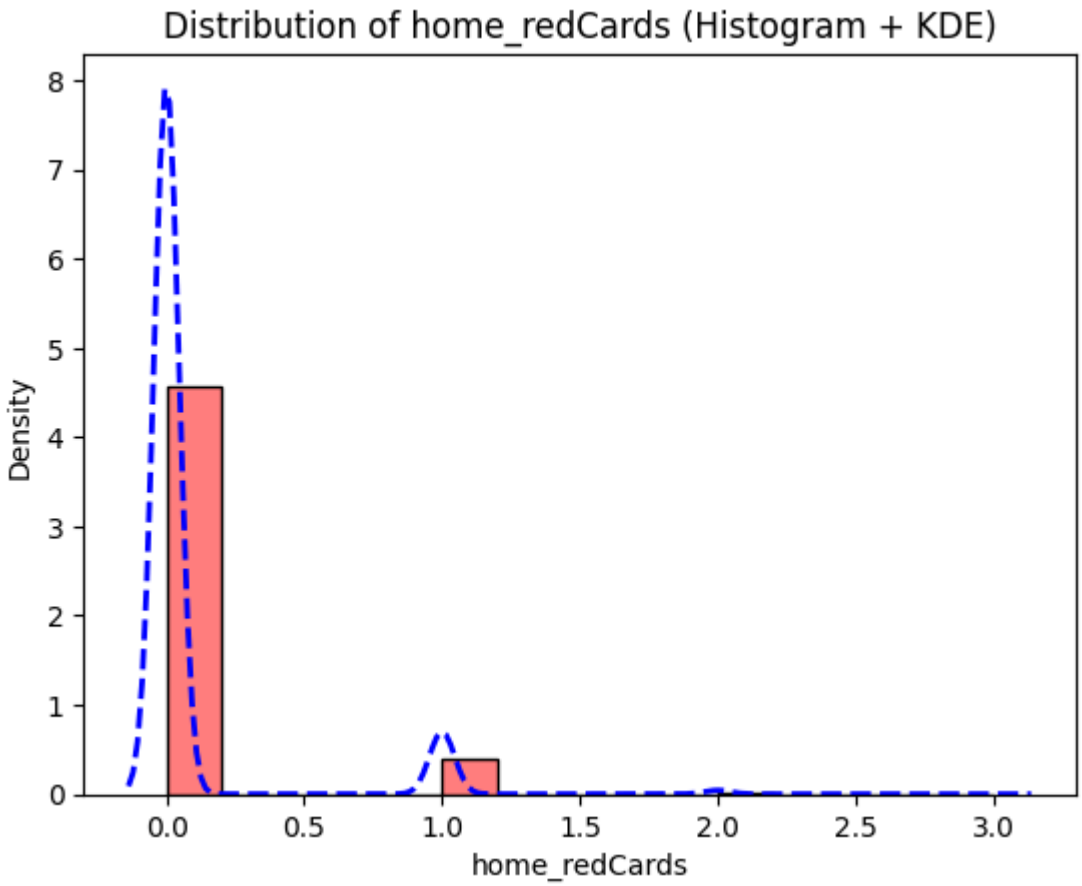

Distribution of home_deep (Histogram + KDE)
Density
0.00
0.05
0.10
0.15
0.20
0
10
20
30
40
home_deep

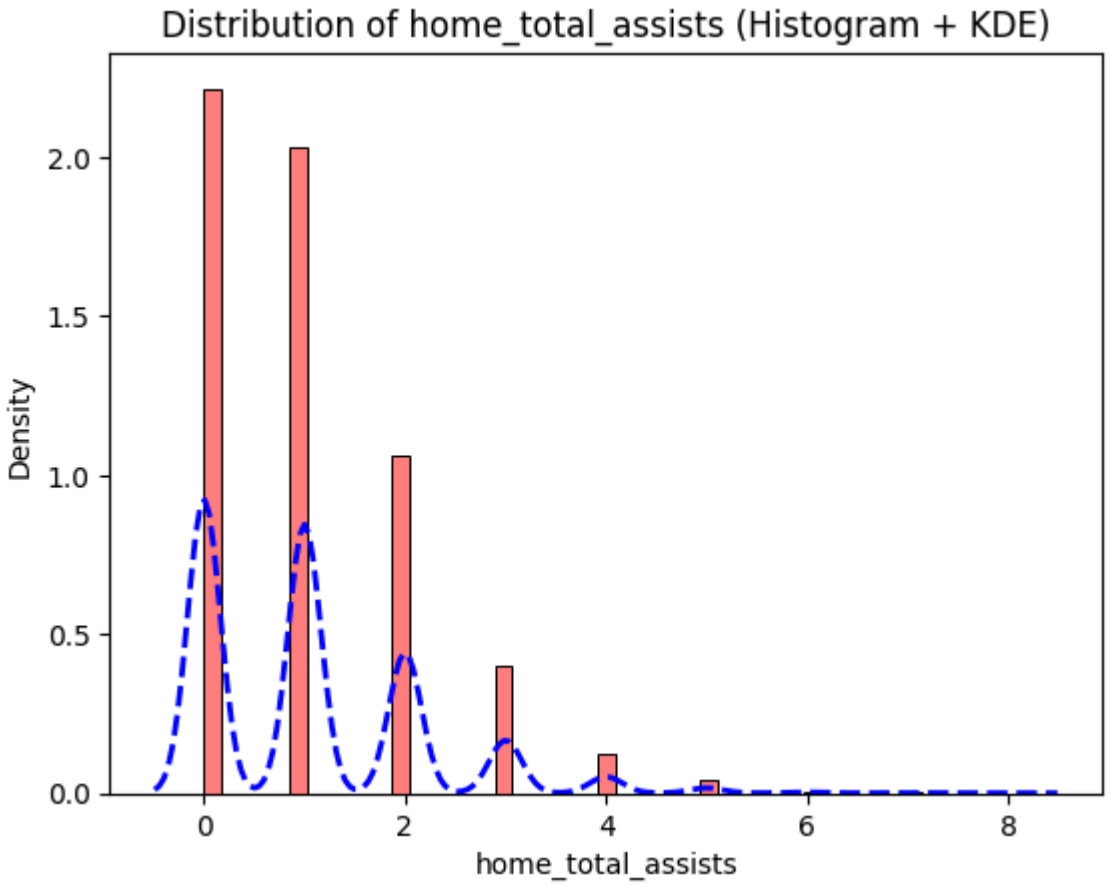

Distribution of home_ppda (Histogram + KDE)
Density
0.00
0.02
0.04
0.06
0.08
0.10
0
20
40
60
80
100
home_ppda

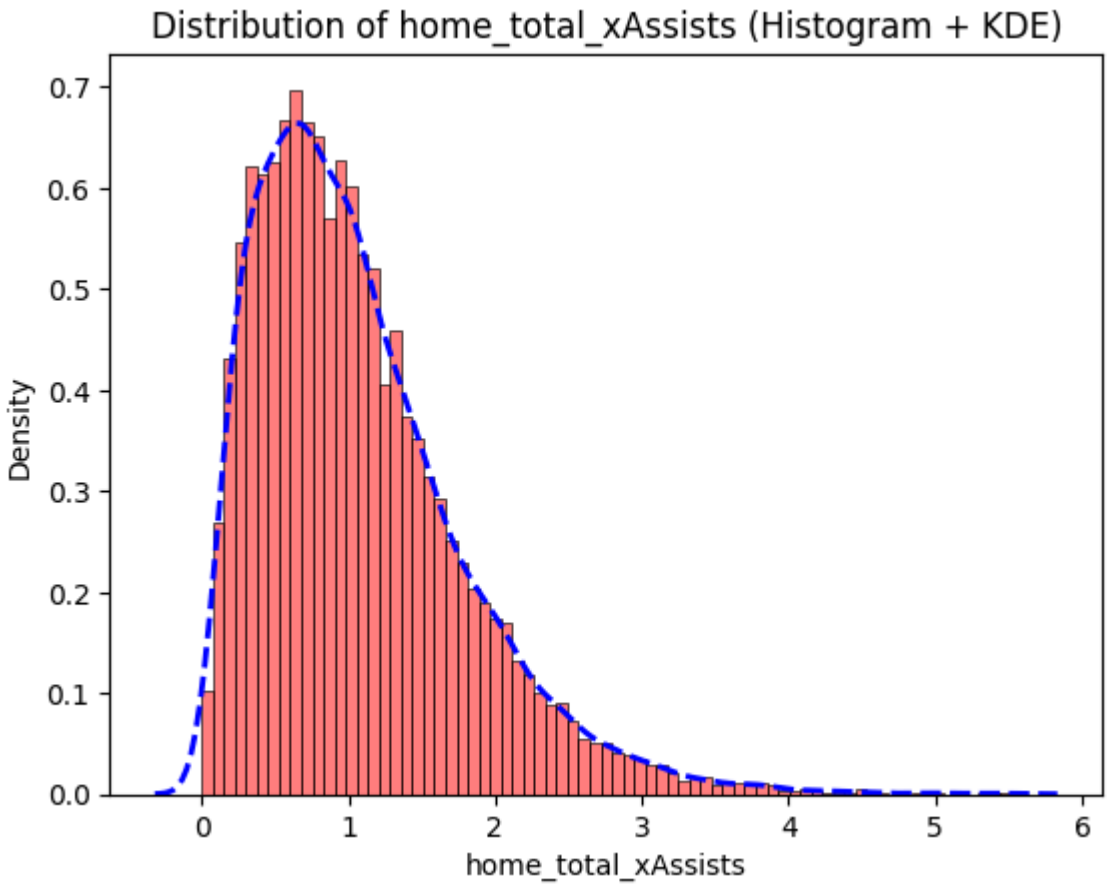

Distribution of home_fouls (Histogram + KDE)
Density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
25
30
35
home_fouls

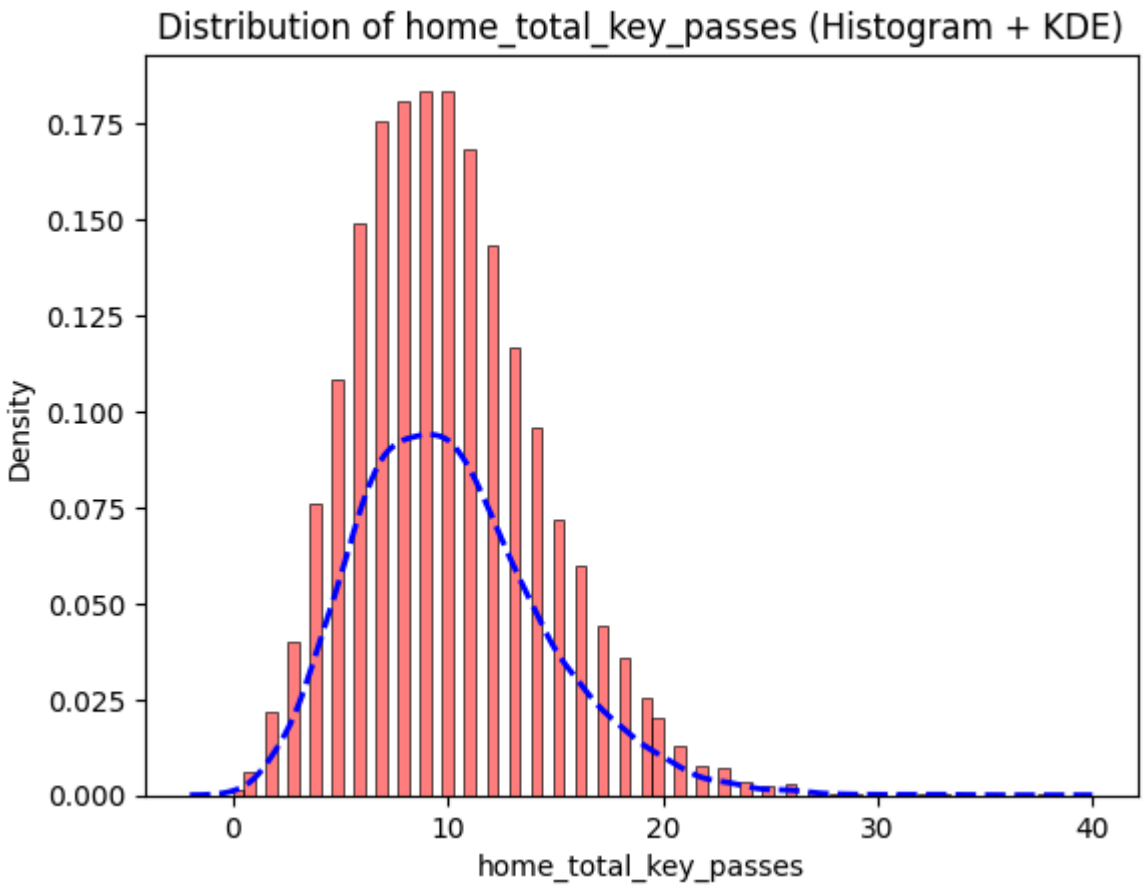

Distribution of home_corners (Histogram + KDE)
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
5
10
15
20
home_corners

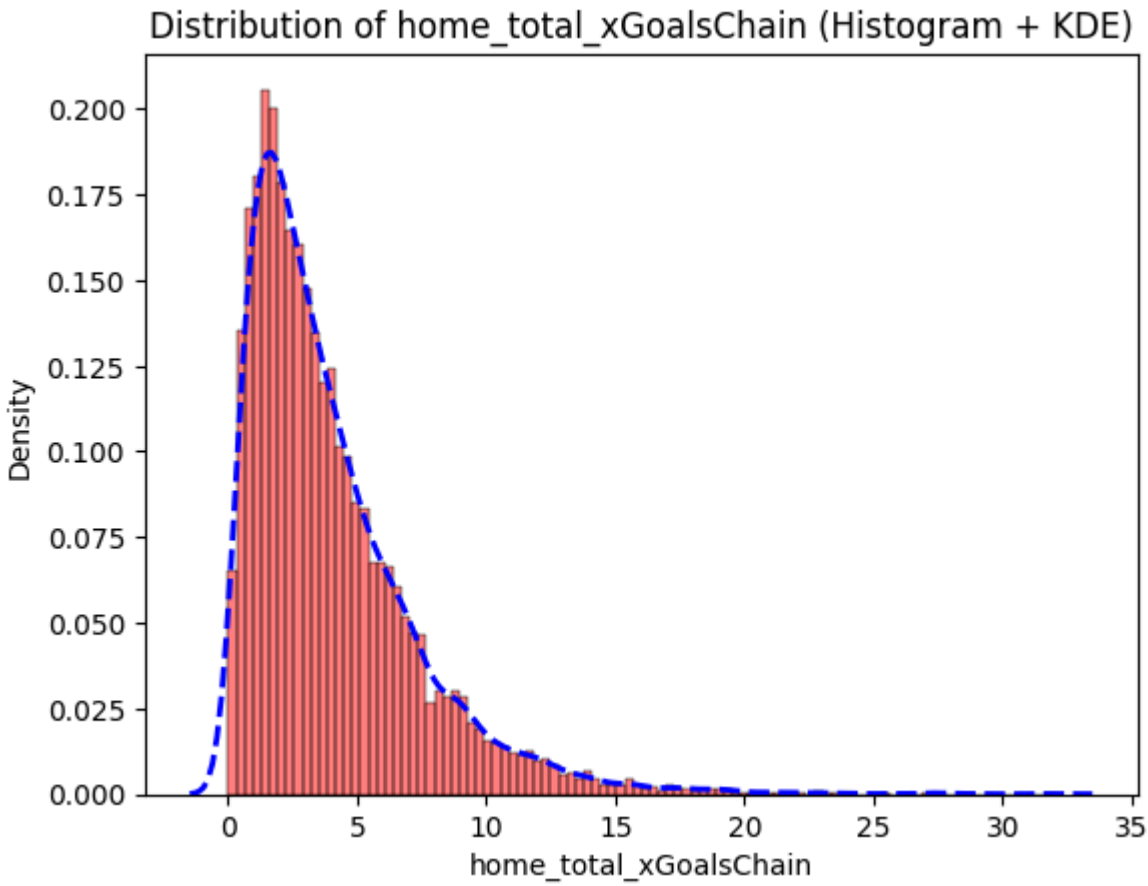

Distribution of home_yellowCards (Histogram + KDE)
Density
home_yellowCards
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0
2
4
6
8

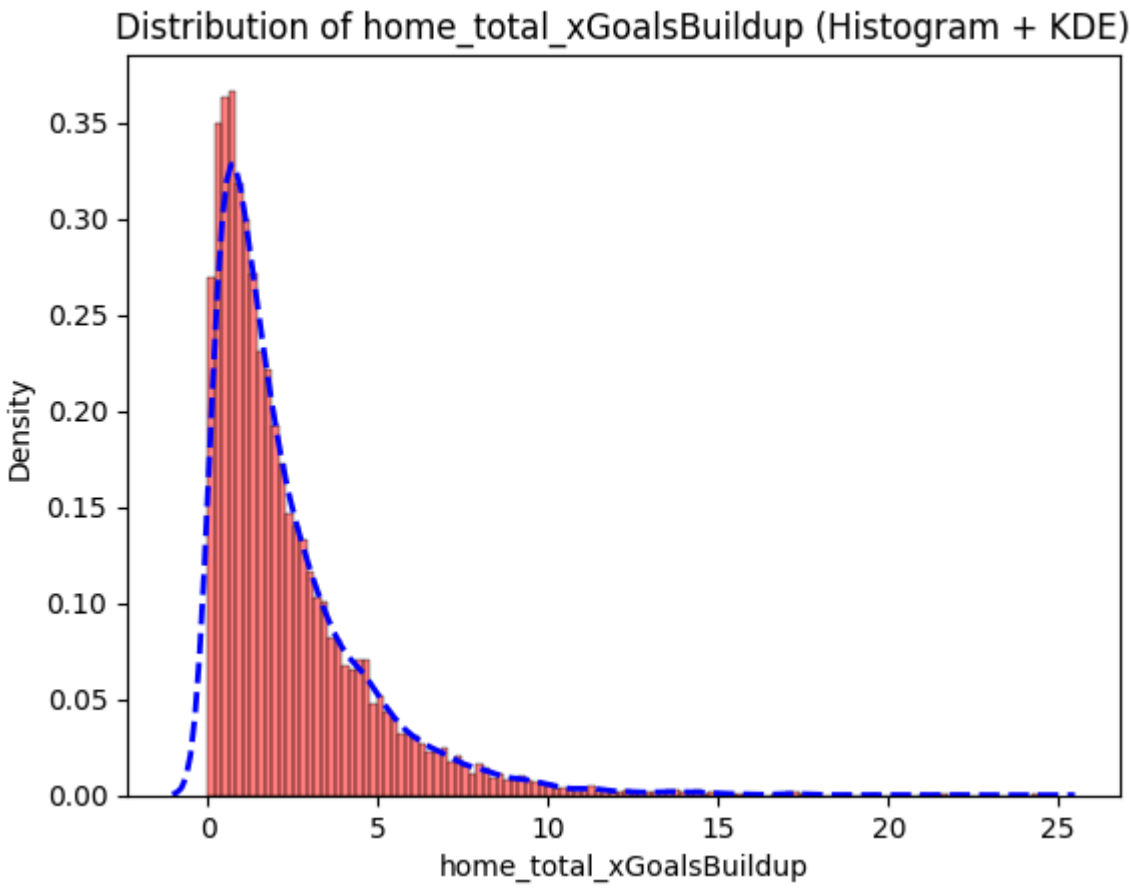

Distribution of home_redCards (Histogram + KDE)
Density
home_redCards
0
1
2
3
4
5
6
7
8
0.0
0.5
1.0
1.5
2.0
2.5
3.0

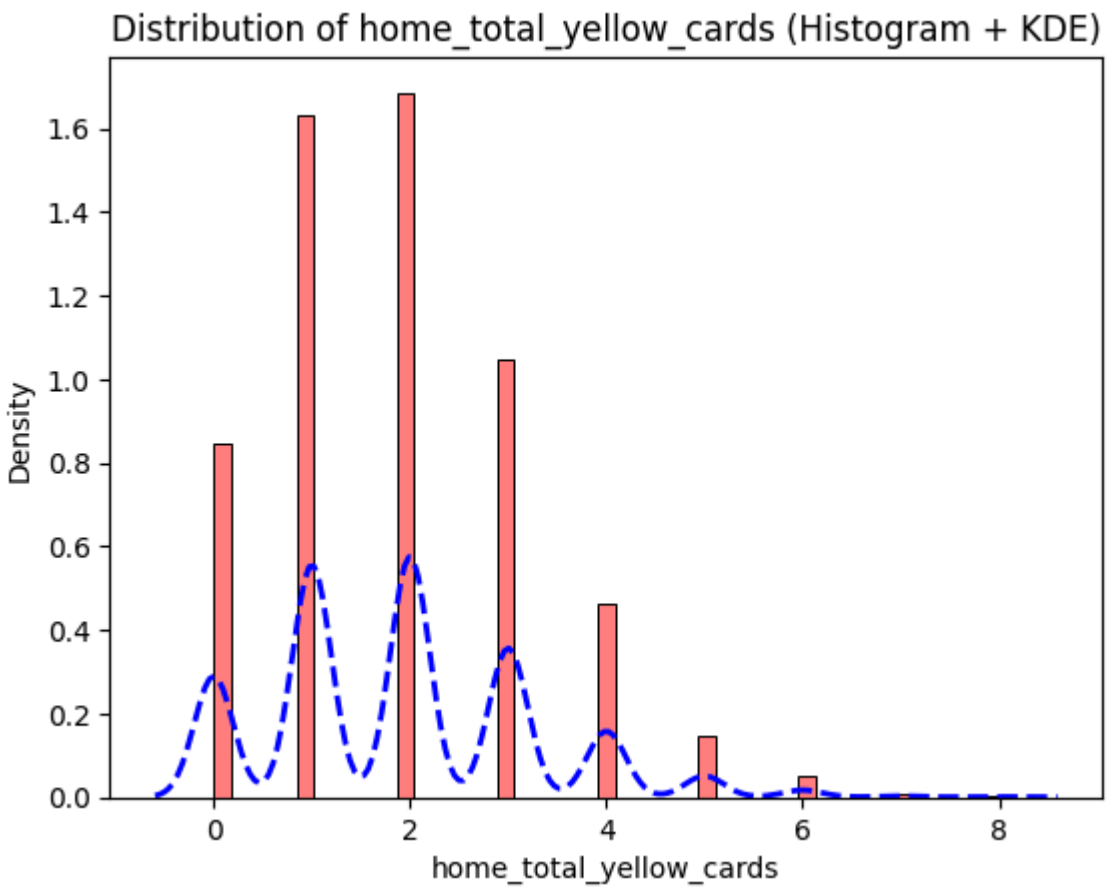

Distribution of home_total_assists (Histogram + KDE)
Density
home_total_assists
0.0
0.5
1.0
1.5
2.0
0
2
4
6
8

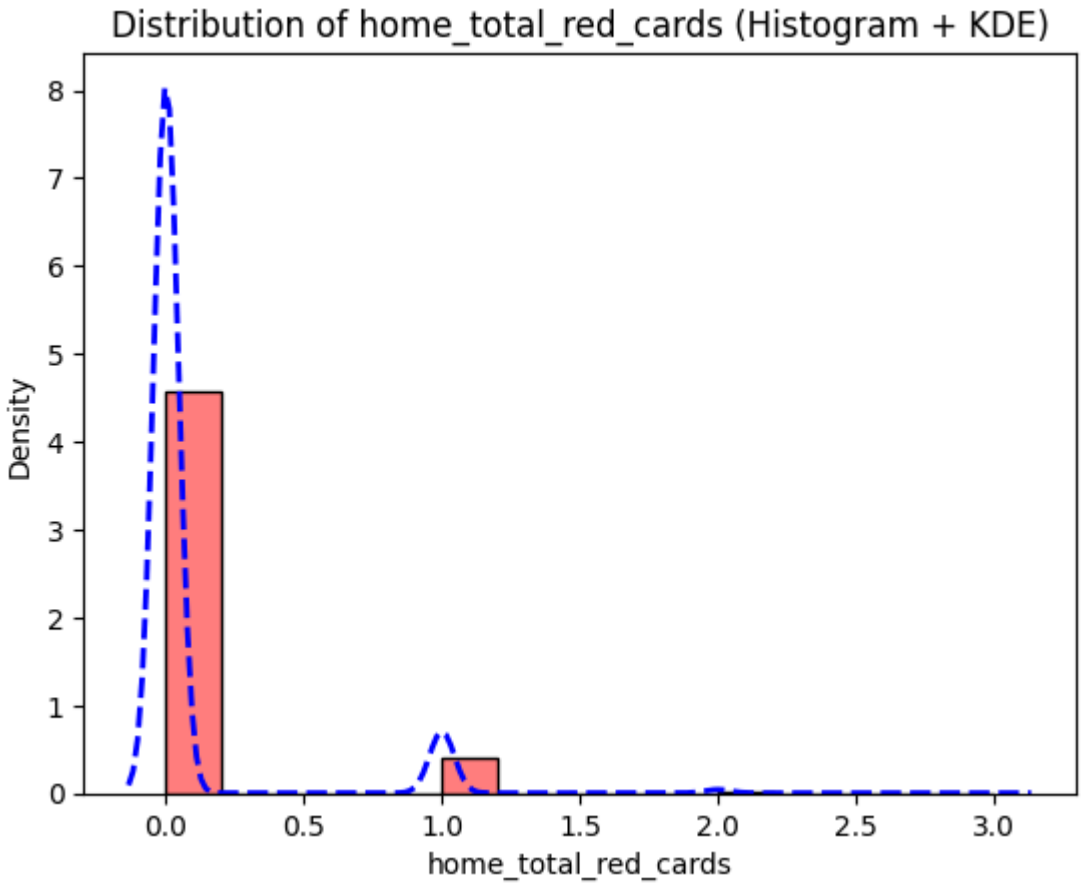

Distribution of home_total_xAssists (Histogram + KDE)
Density
home_total_xAssists
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
2
3
4
5
6

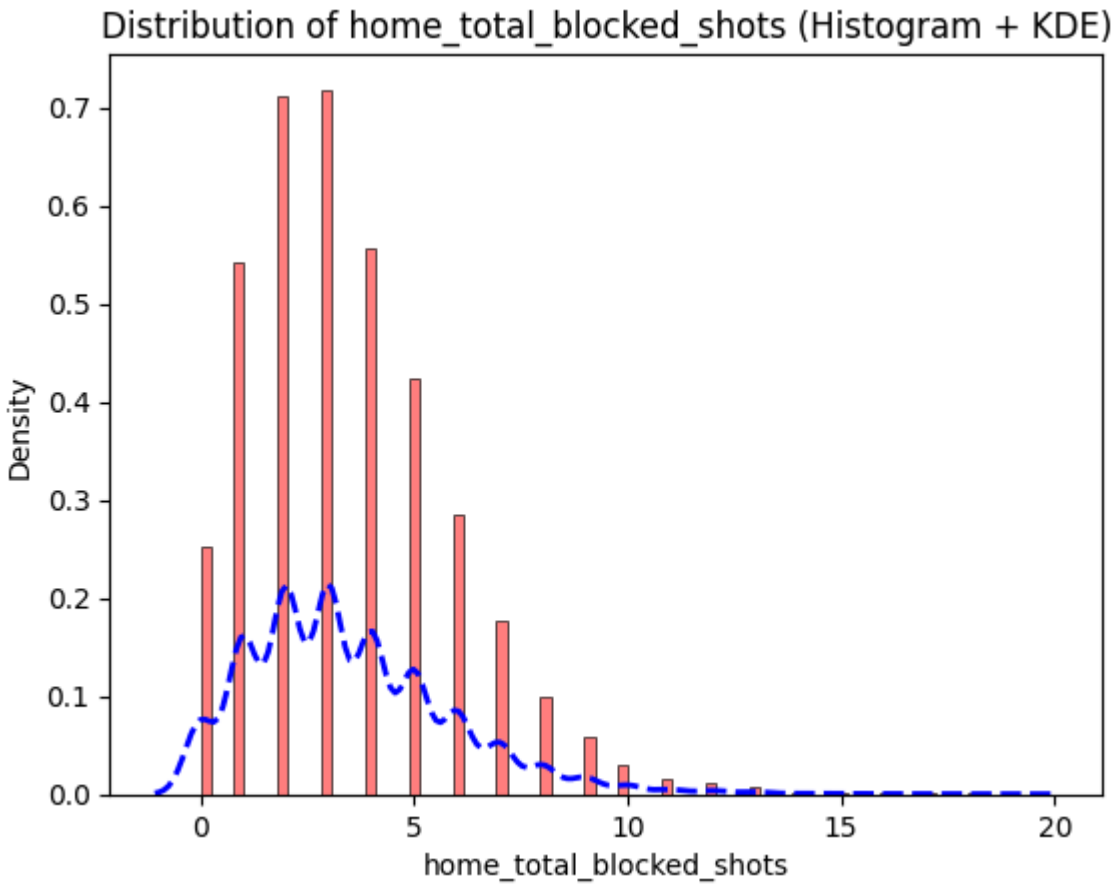

Distribution of home_total_key_passes (Histogram + KDE)
Density
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
10
20
30
40
home_total_key_passes

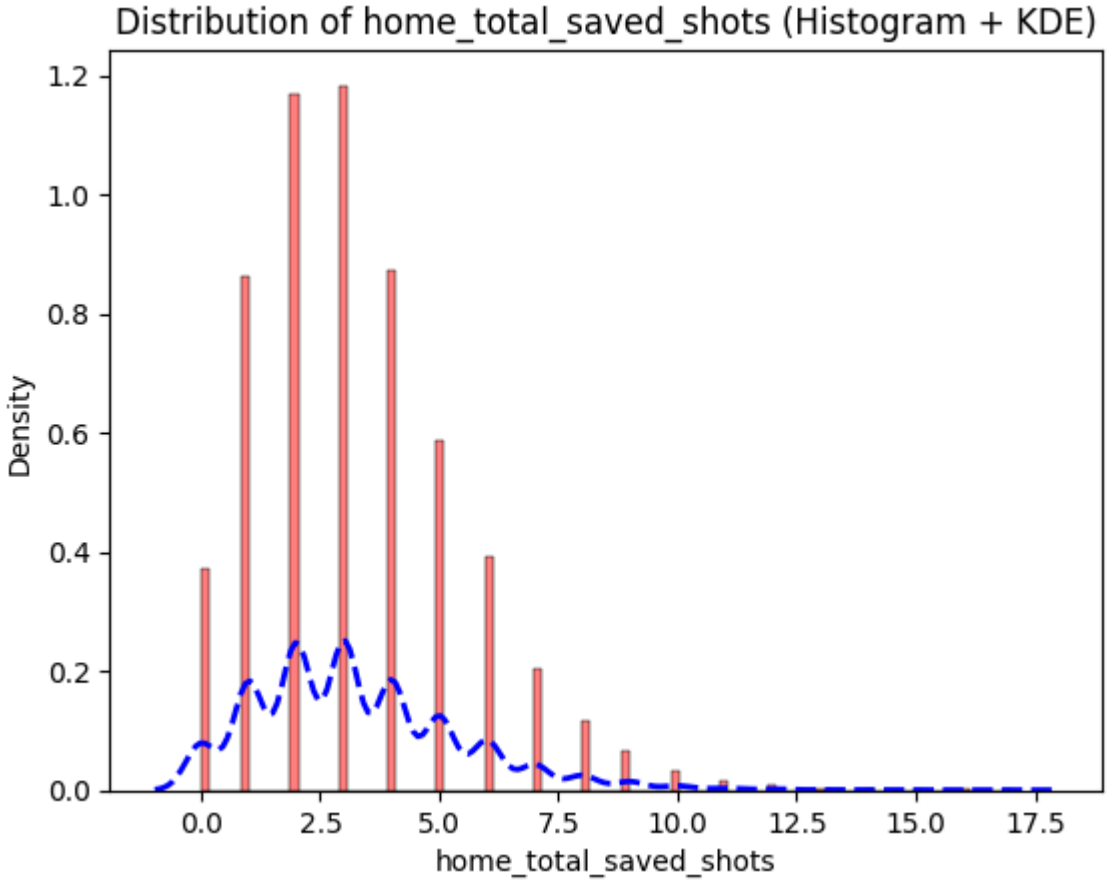

Distribution of home_total_xGoalsChain (Histogram + KDE)
Density
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
5
10
15
20
25
30
35
home_total_xGoalsChain

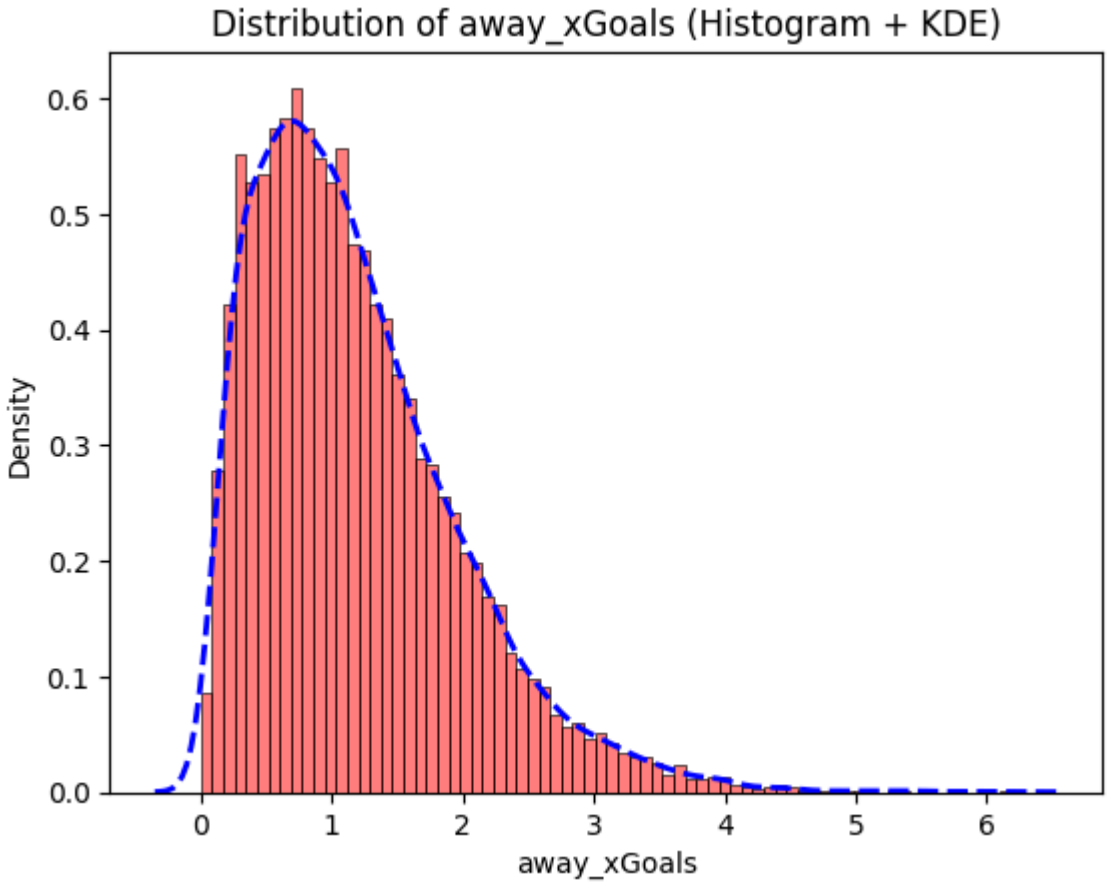

Distribution of home_total_xGoalsBuildup (Histogram + KDE)
Density
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
25
home_total_xGoalsBuildup

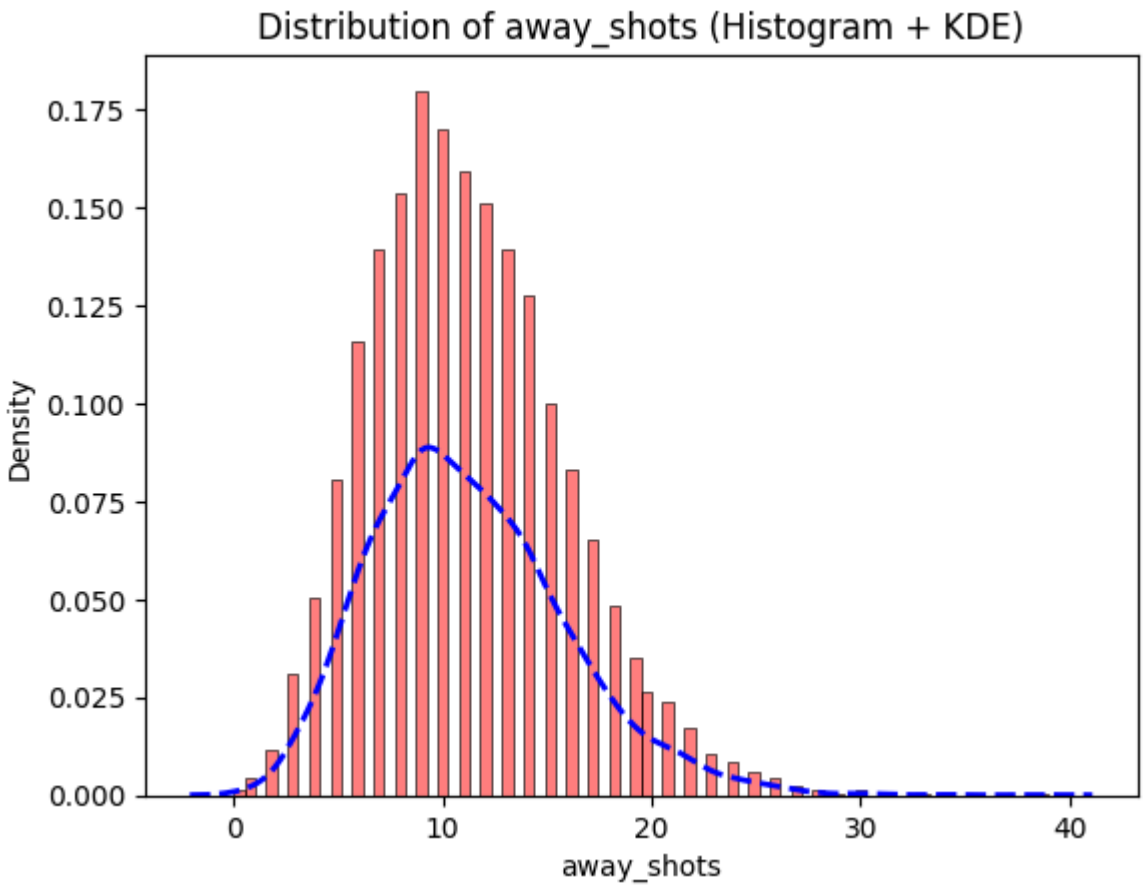

Distribution of home_total_yellow_cards (Histogram + KDE)
Density
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
home_total_yellow_cards

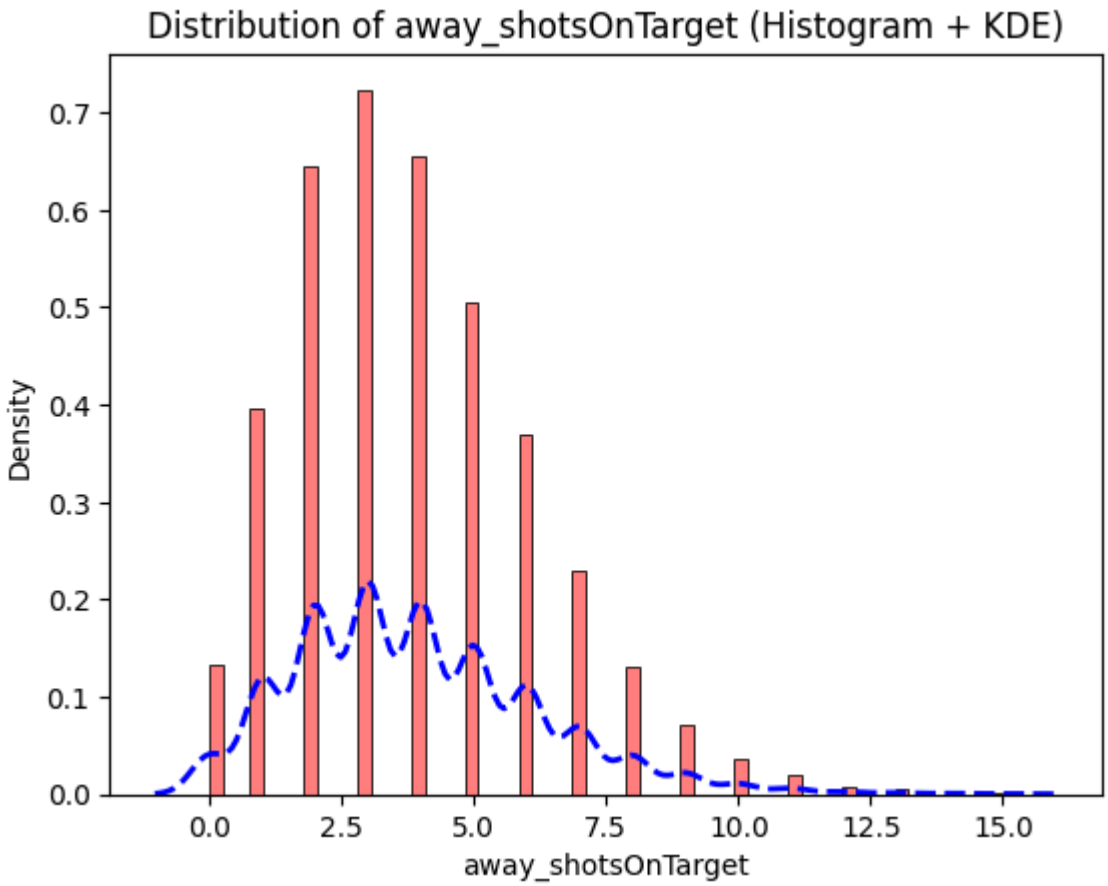

Distribution of home_total_red_cards (Histogram + KDE)
Density
home_total_red_cards
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
6
7
8

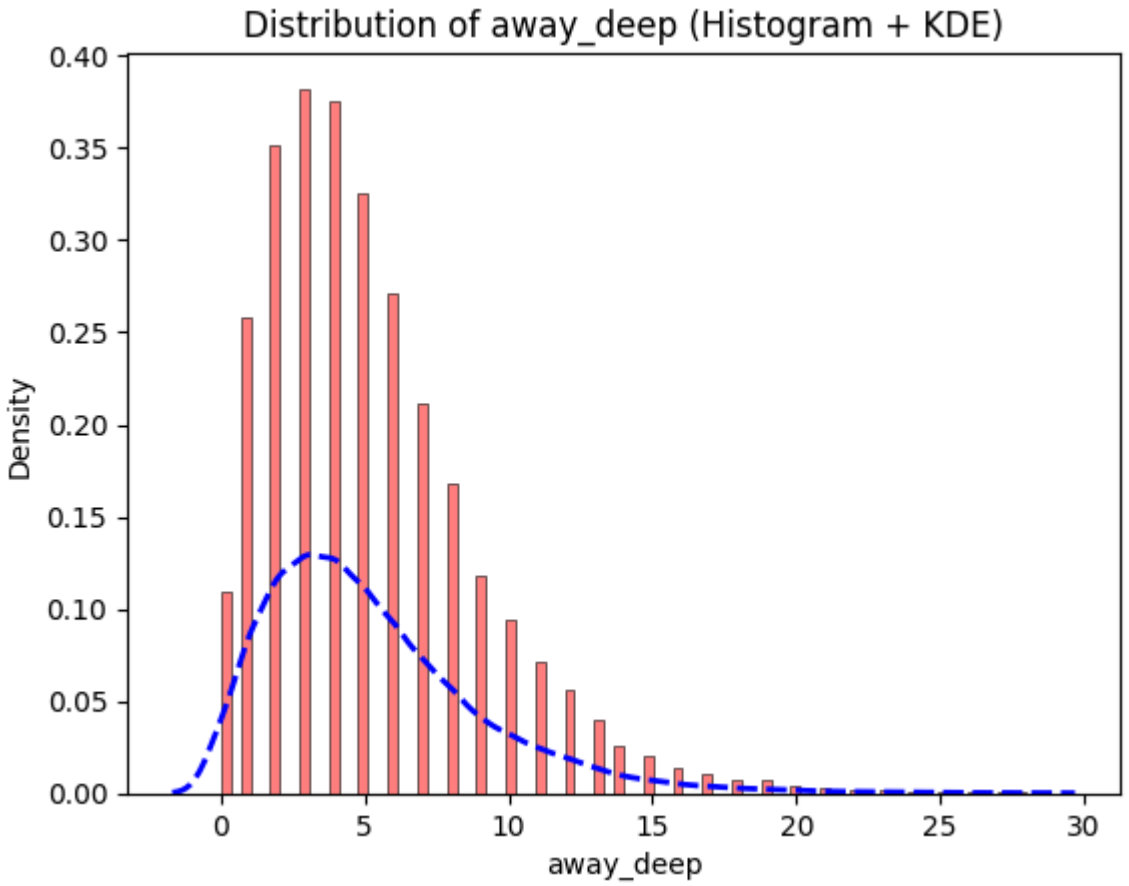

Distribution of home_total_blocked_shots (Histogram + KDE)
Density
home_total_blocked_shots
0
5
10
15
20
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

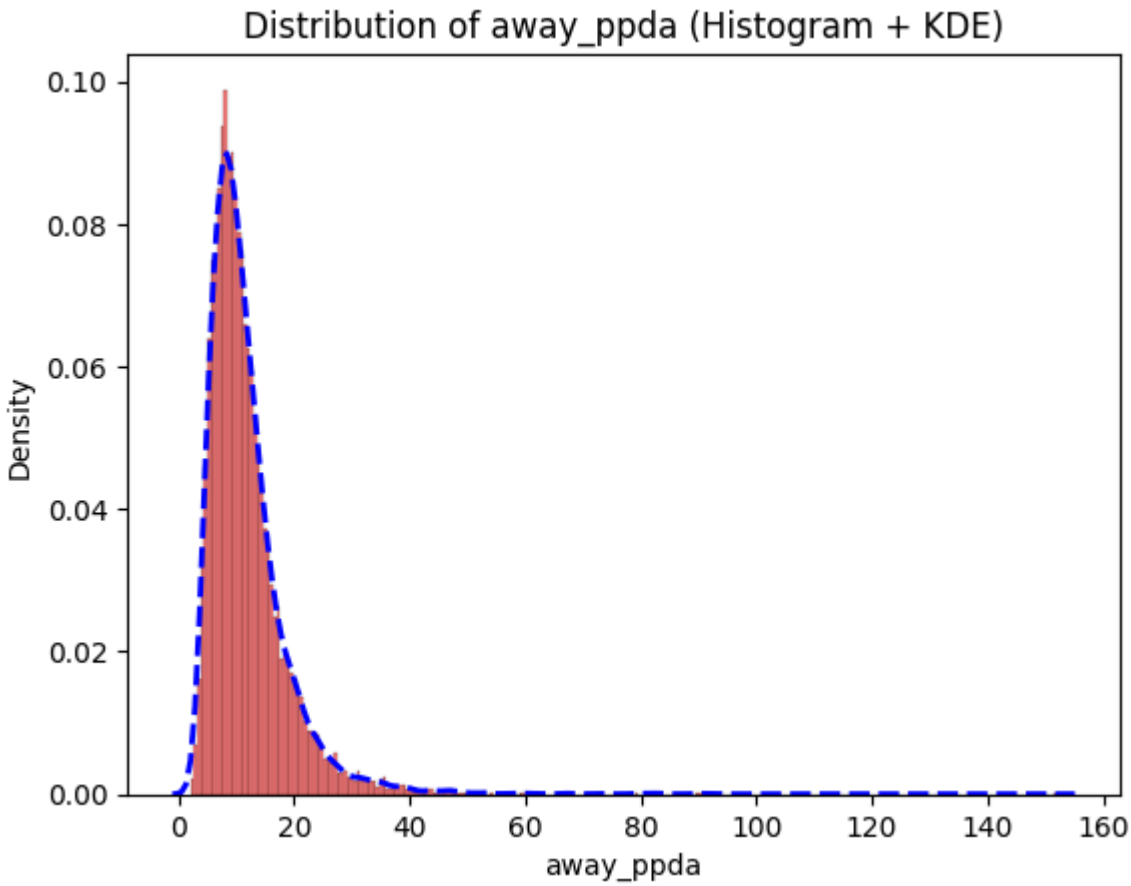

Distribution of home_total_saved_shots (Histogram + KDE)
Density
home_total_saved_shots
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

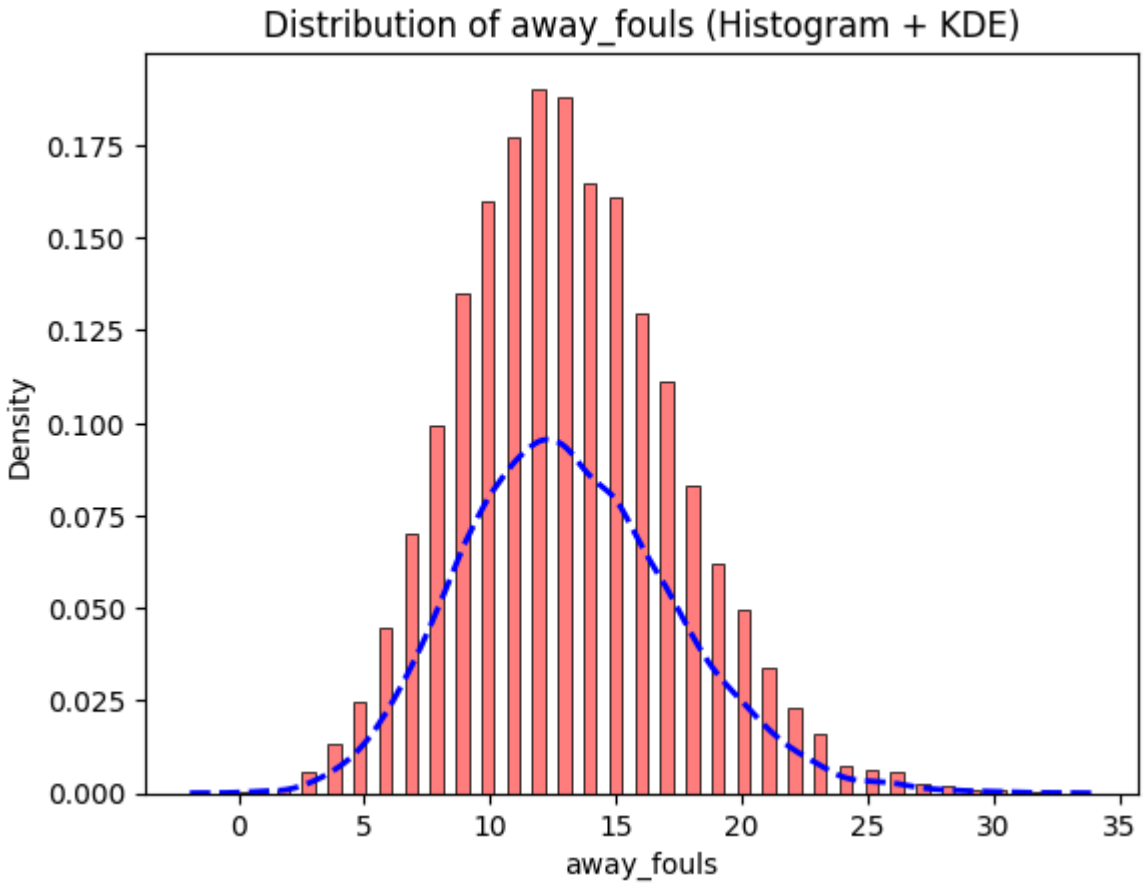

Distribution of away_xGoals (Histogram + KDE)
Density
away_xGoals
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
2
3
4
5
6

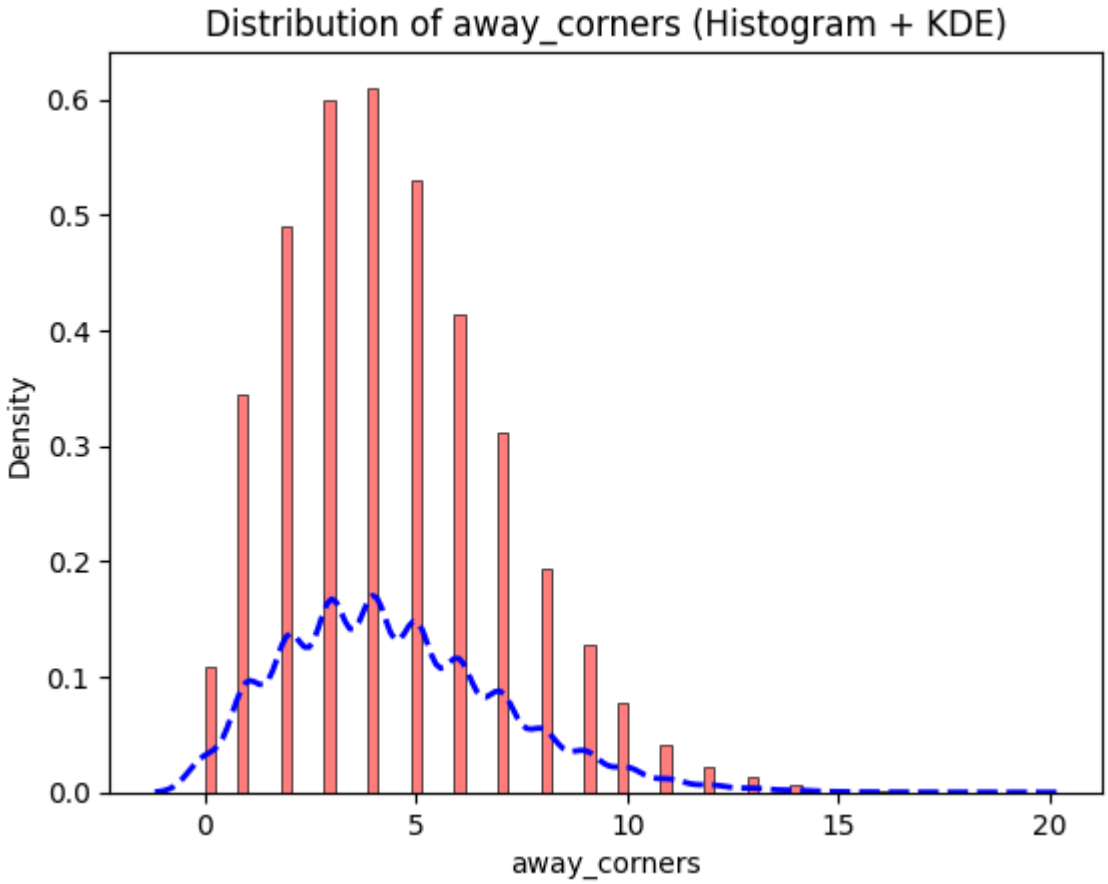

Distribution of away_shots (Histogram + KDE)
Density
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
10
20
30
40
away_shots

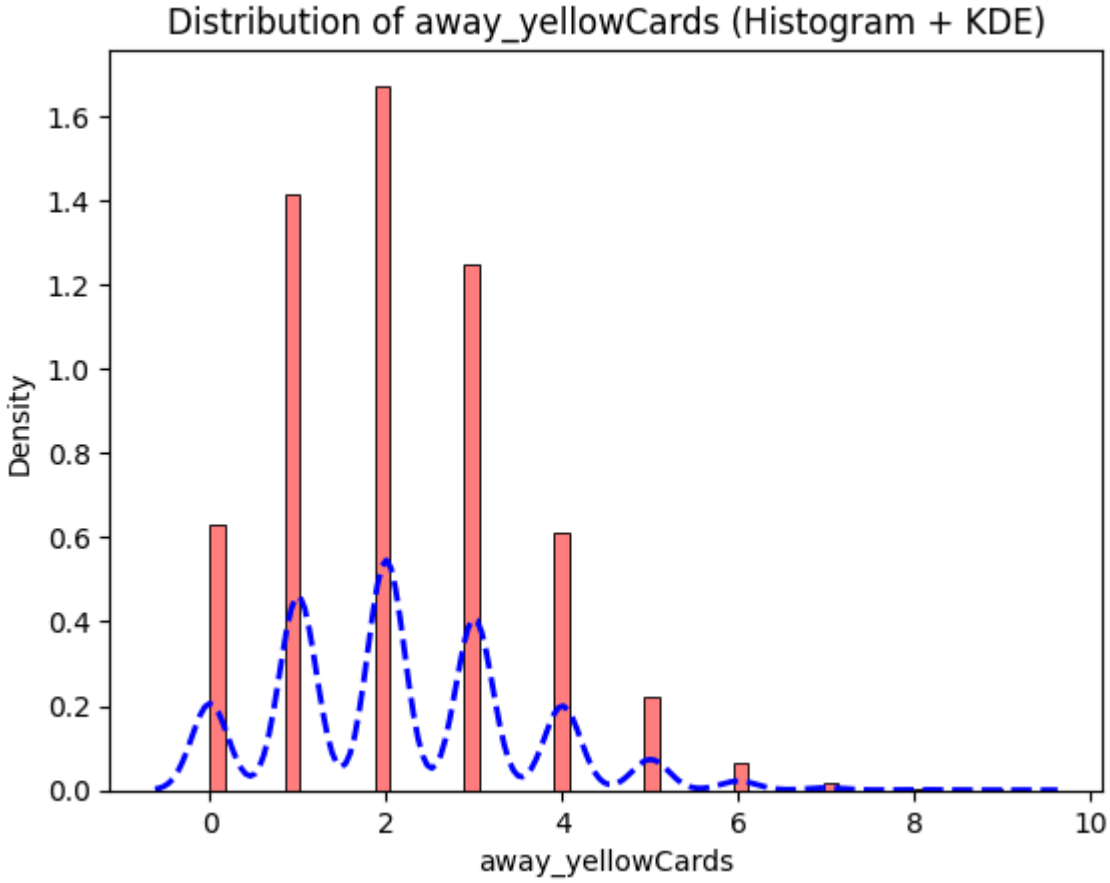

Distribution of away_shotsOnTarget (Histogram + KDE)
Density
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
away_shotsOnTarget

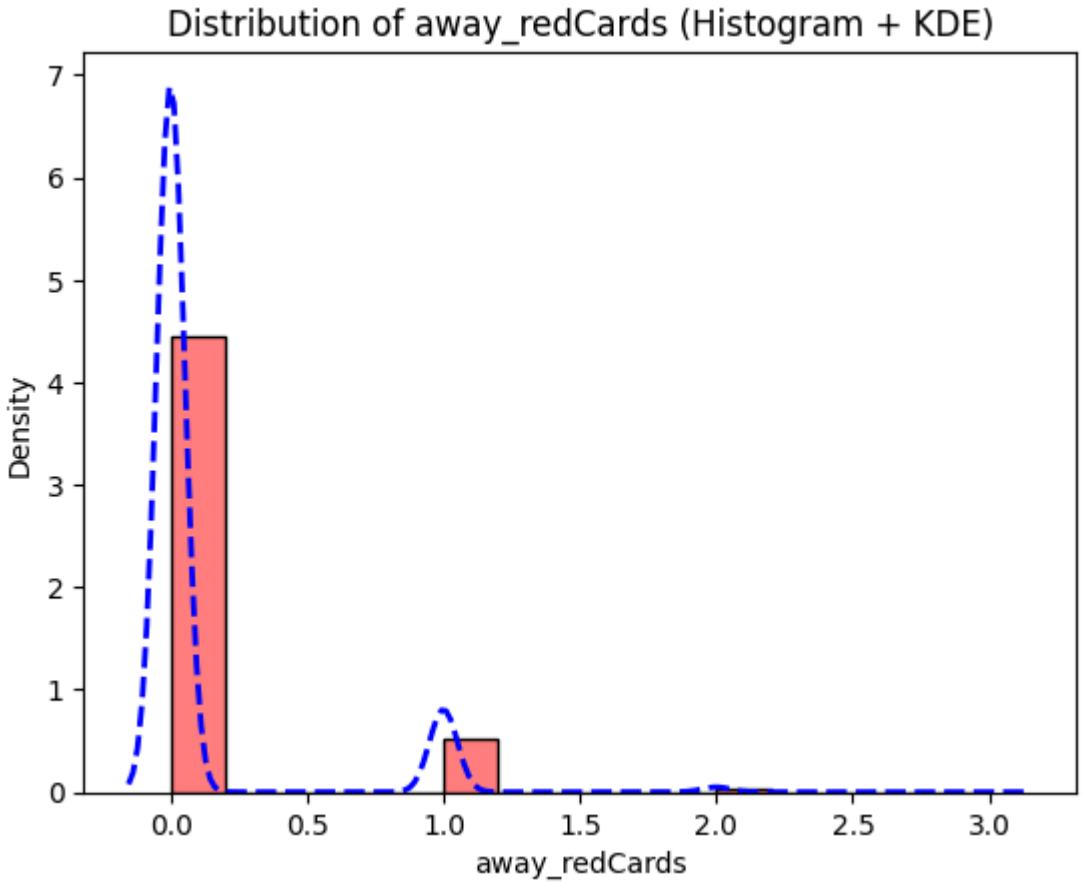

Distribution of away_deep (Histogram + KDE)
Density
away_deep
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
5
10
15
20
25
30

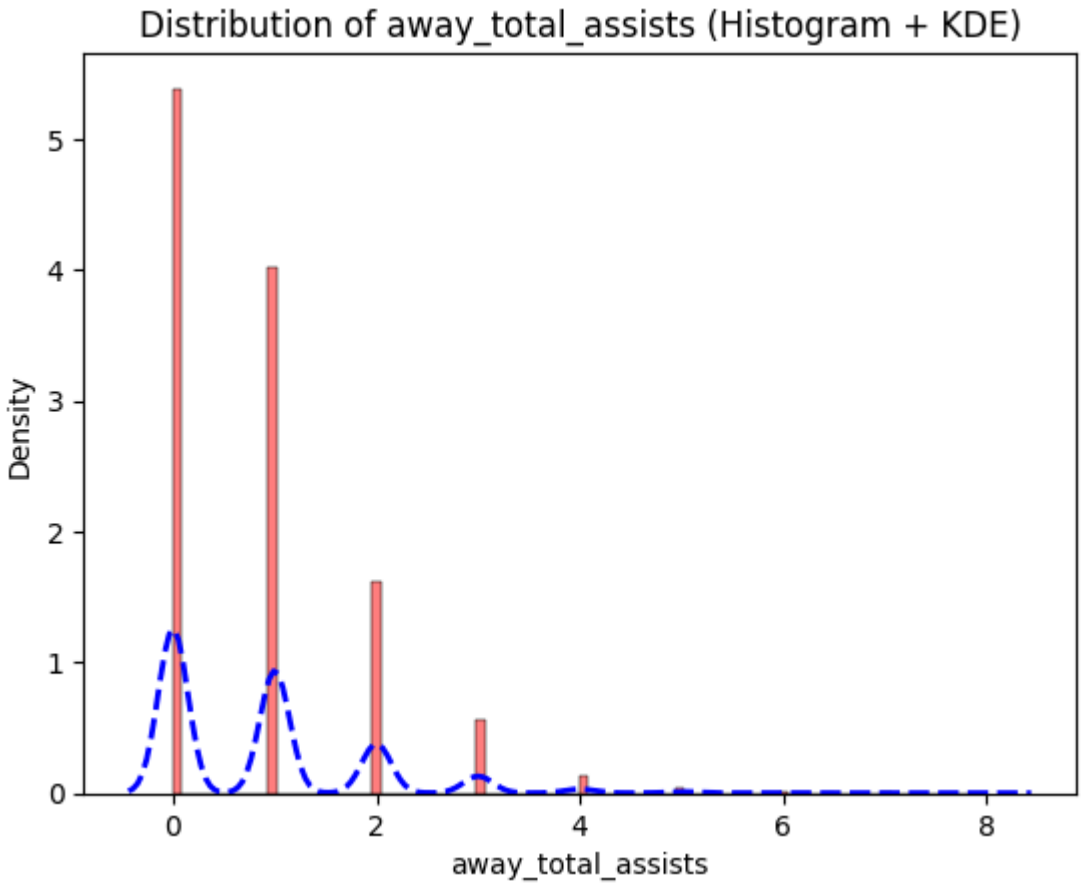

Distribution of away_ppda (Histogram + KDE)
Density
away_ppda
0.00
0.02
0.04
0.06
0.08
0.10
0
20
40
60
80
100
120
140
160

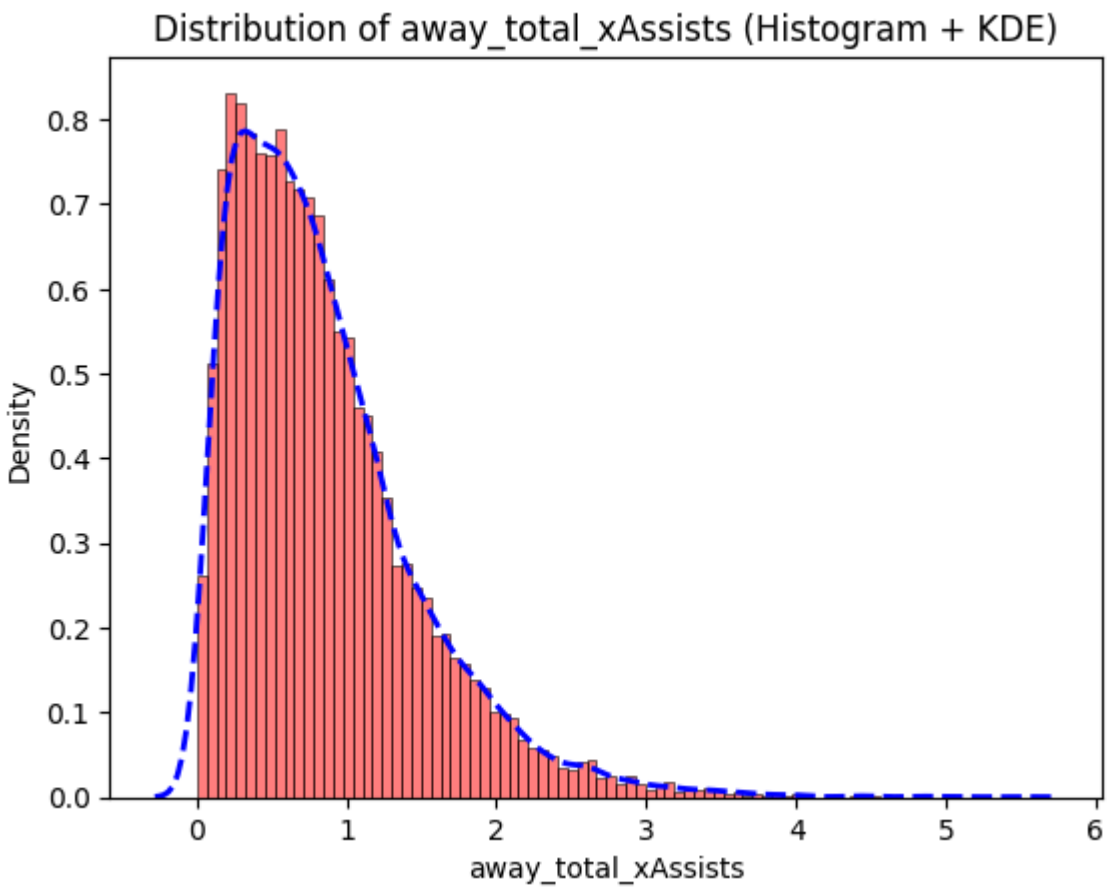

Distribution of away_fouls (Histogram + KDE)
Density
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
5
10
15
20
25
30
35
away_fouls

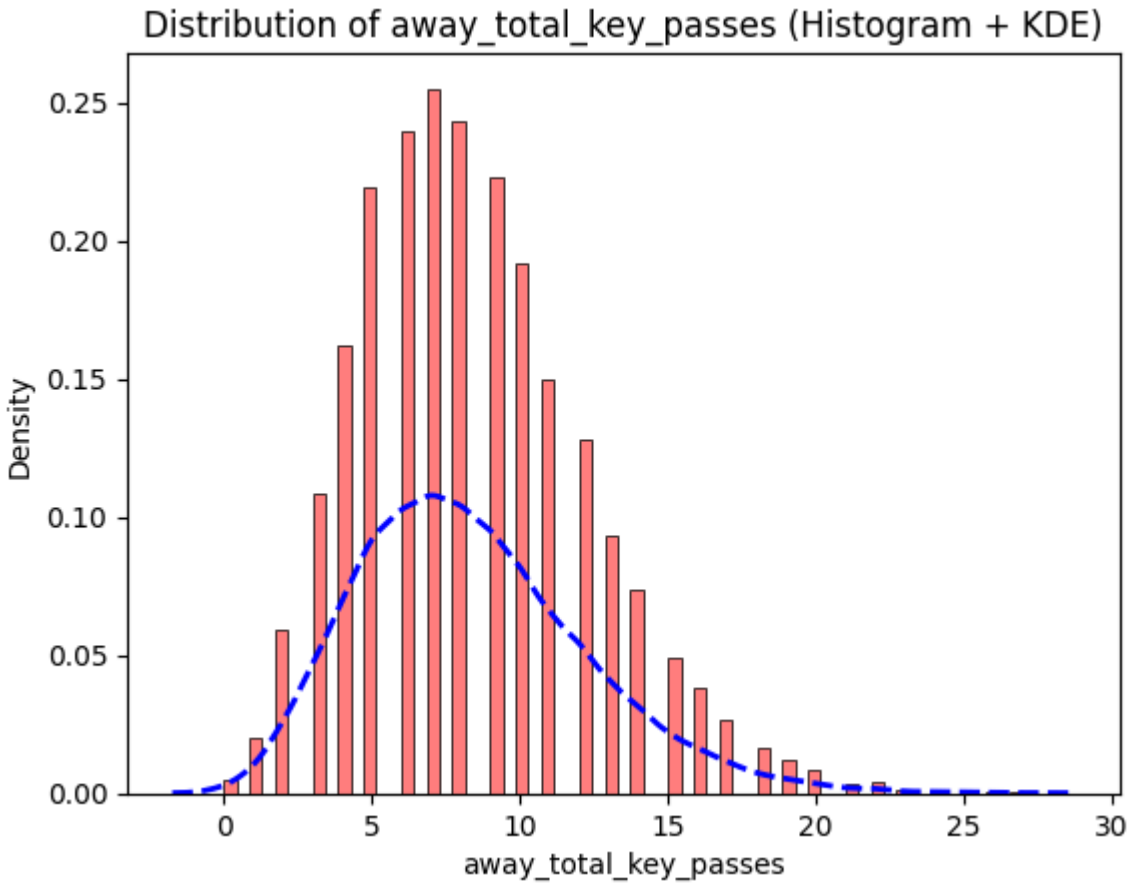

Distribution of away_corners (Histogram + KDE)
Density
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
away_corners

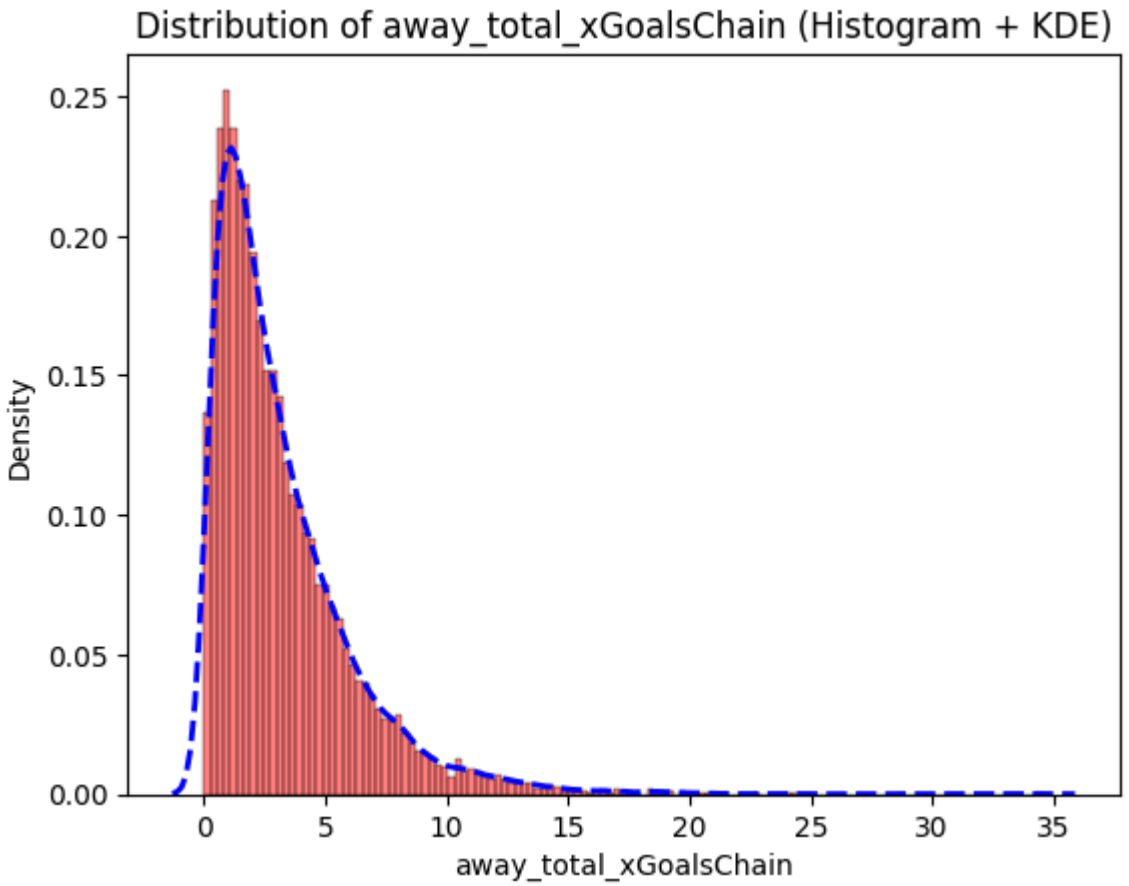

Distribution of away_yellowCards (Histogram + KDE)
Density
away_yellowCards
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0
2
4
6
8
10

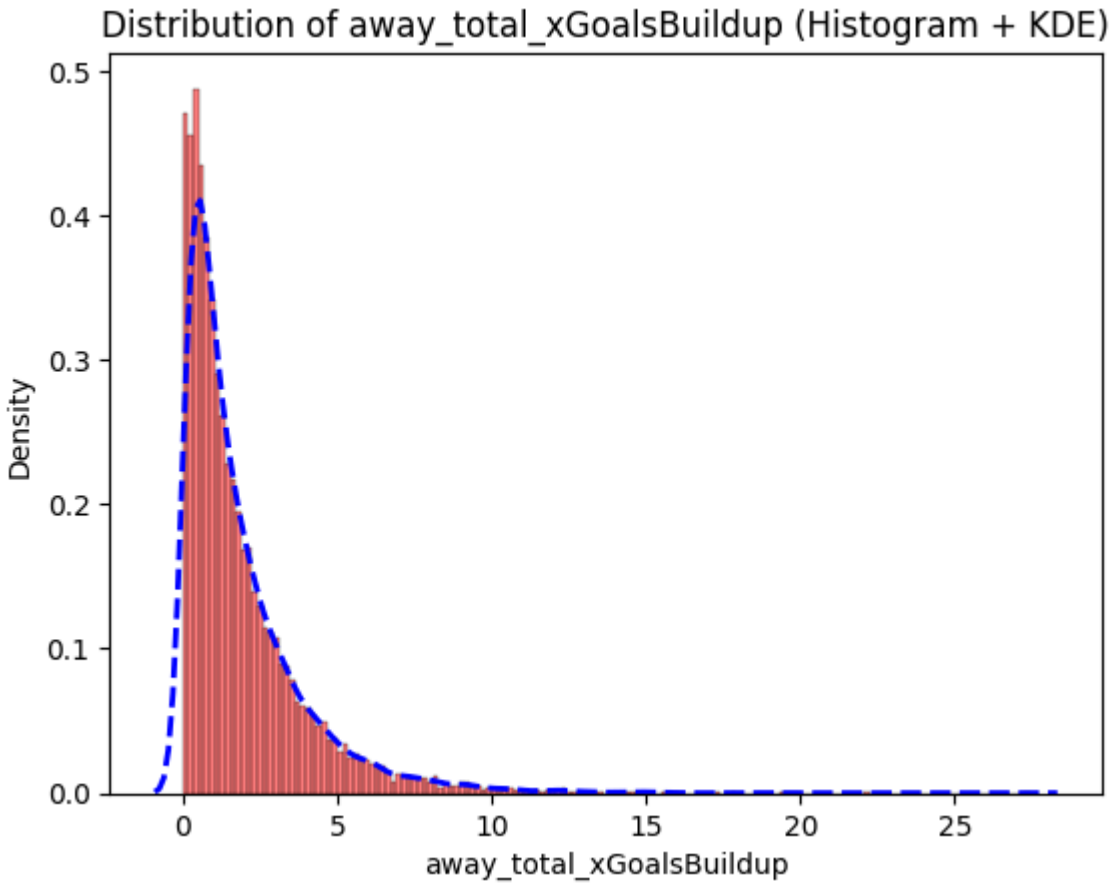

Distribution of away_redCards (Histogram + KDE)
Density
away_redCards
0
1
2
3
4
5
6
7
0.0
0.5
1.0
1.5
2.0
2.5
3.0

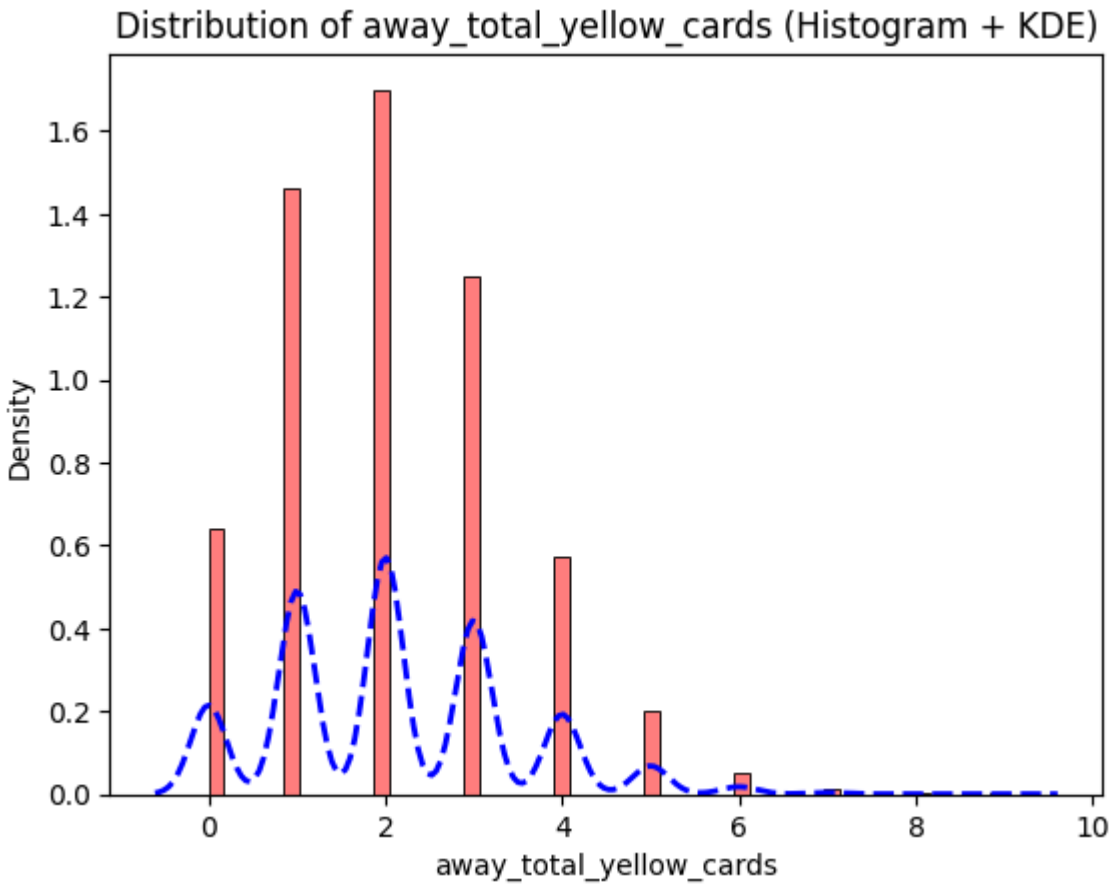

Distribution of away_total_assists (Histogram + KDE)
Density
away_total_assists
0
1
2
3
4
5
0
2
4
6
8

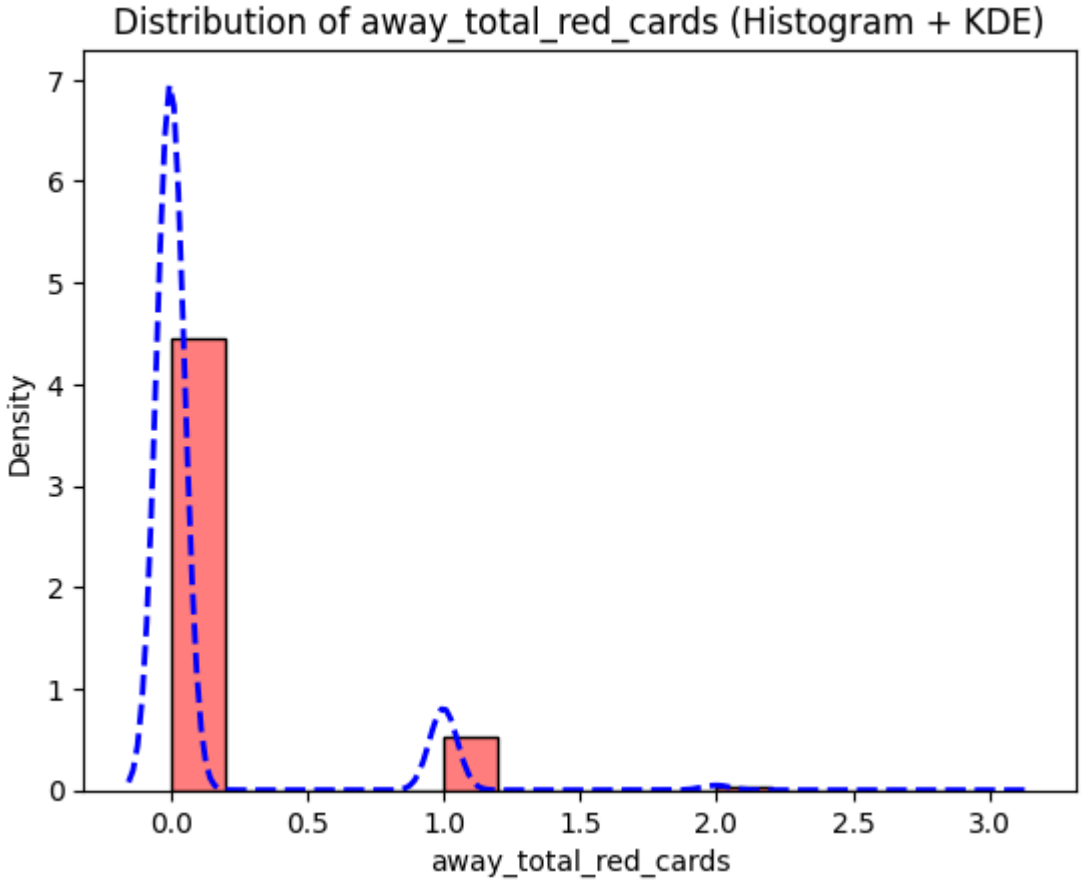

Distribution of away_total_xAssists (Histogram + KDE)
Density
away_total_xAssists
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
1
2
3
4
5
6

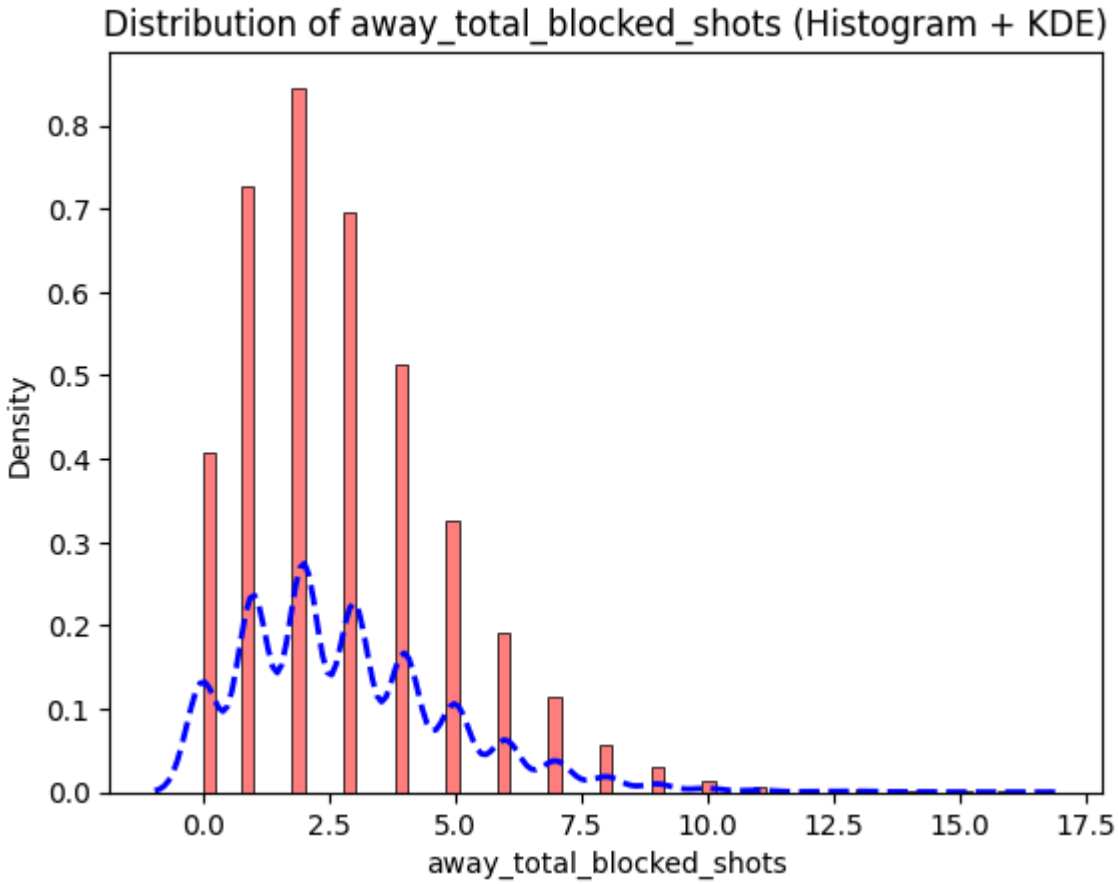

Distribution of away_total_key_passes (Histogram + KDE)
Density
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
25
30
away_total_key_passes

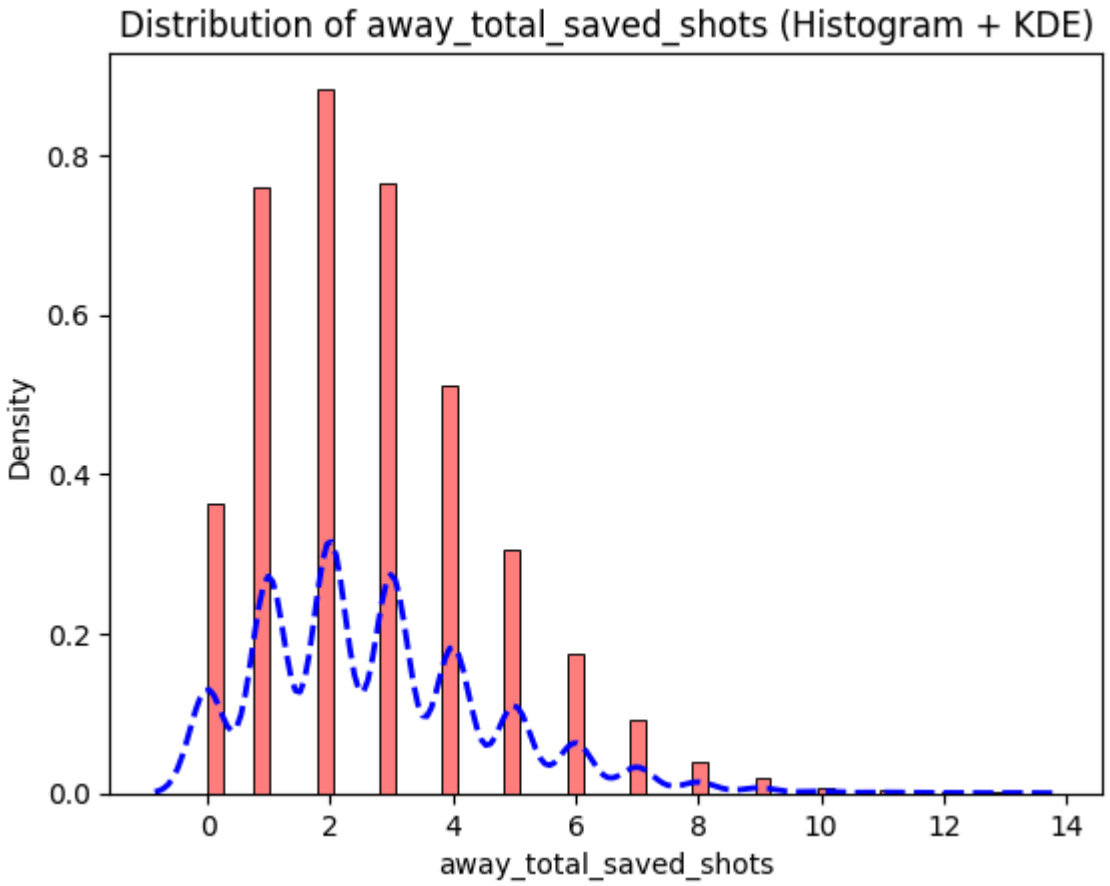

Distribution of away_total_xGoalsChain (Histogram + KDE)
Density
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
25
30
35
away_total_xGoalsChain


Distribution of away_total_xGoalsBuildup (Histogram + KDE)
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
20
25
away_total_xGoalsBuildup


Distribution of away_total_yellow_cards (Histogram + KDE)
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0
2
4
6
8
10
away_total_yellow_cards


Distribution of away_total_red_cards (Histogram + KDE)
Density
away_total_red_cards
0
1
2
3
4
5
6
7
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Distribution of away_total_blocked_shots (Histogram + KDE)
Density
away_total_blocked_shots
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5



DateTime

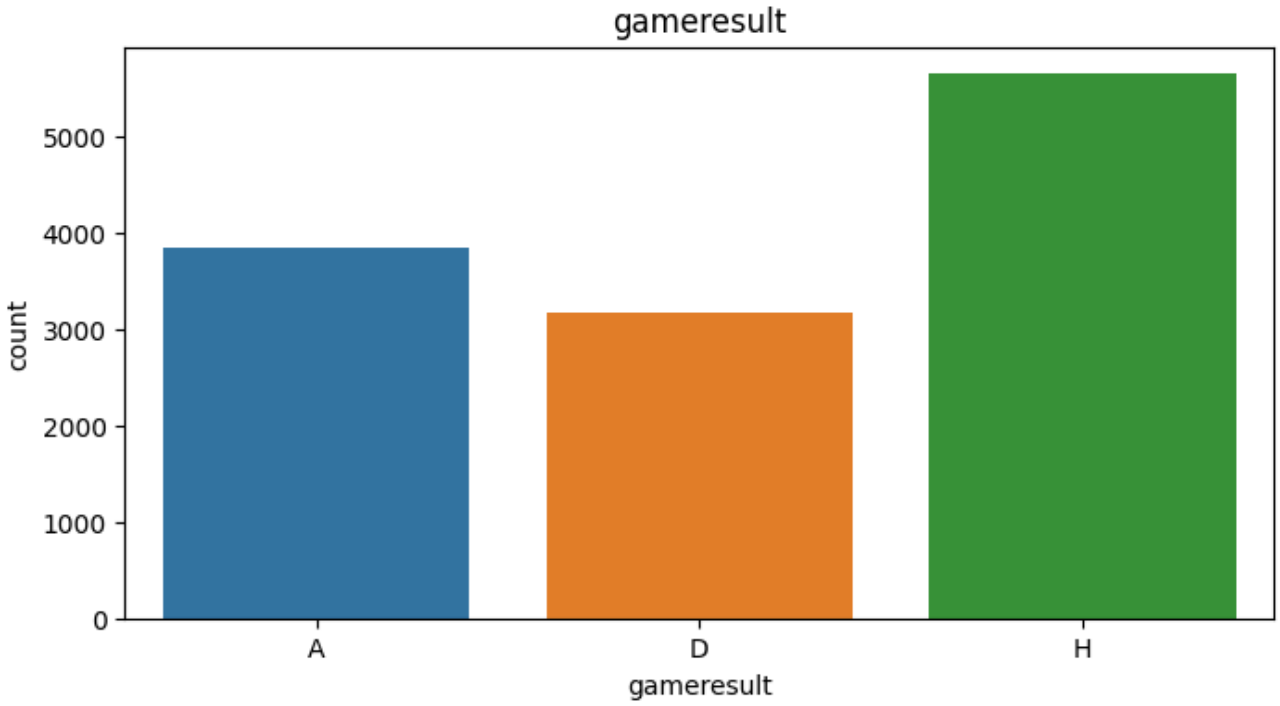
```
0       2015-08-08 15:45:00
1       2015-08-08 18:00:00
2       2015-08-08 18:00:00
3       2015-08-08 18:00:00
4       2015-08-08 18:00:00
               ...
12675   2021-05-23 19:00:00
12676   2021-05-23 19:00:00
12677   2021-05-23 19:00:00
12678   2021-05-23 19:00:00
12679   2021-05-23 19:00:00
Name: date, Length: 12680, dtype: datetime64[ns]
```

Categorials

```
0        H
1        A
2        D
3        H
4        A
        ..
12675    A
12676    A
12677    H
12678    A
12679    D
Name: gameresult, Length: 12680, dtype: object
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12680 entries, 0 to 12679
Data columns (total 1 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   gameresult  12680 non-null  category
dtypes: category(1)
memory usage: 111.6 KB
```

```
<Figure size 2500x1000 with 0 Axes>
```

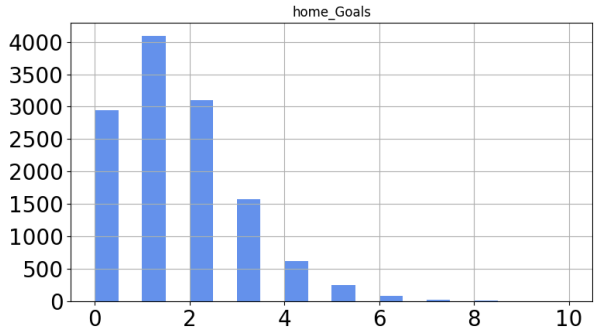


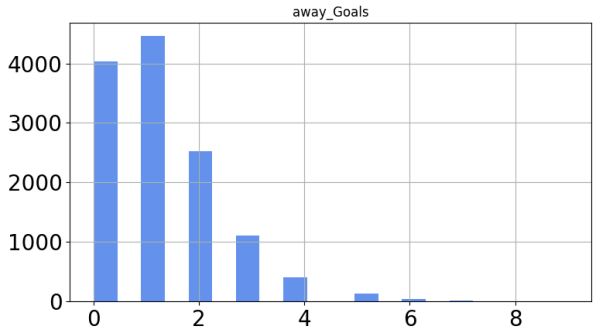
Continues (numeric)

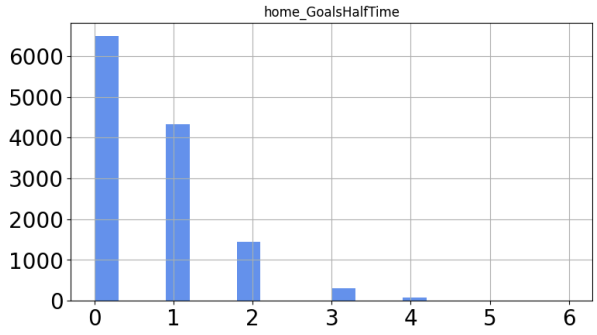
```
(12680, 40)
```

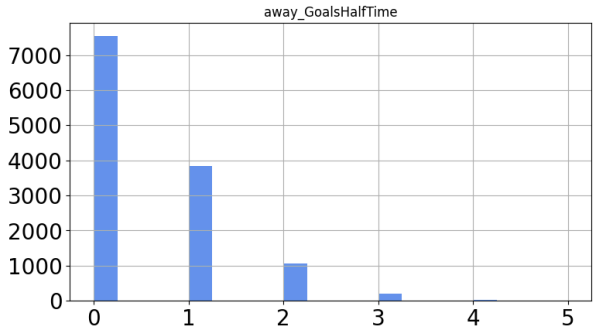
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12680 entries, 0 to 12679
Data columns (total 40 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   home_Goals                12680 non-null  int64  
 1   away_Goals                12680 non-null  int64  
 2   home_GoalsHalfTime        12680 non-null  int64  
 3   away_GoalsHalfTime        12680 non-null  int64  
 4   home_xGoals               12680 non-null  float64
 5   home_shots                12680 non-null  int64  
 6   home_shotsOnTarget        12680 non-null  int64  
 7   home_deep                 12680 non-null  int64  
 8   home_ppda                 12680 non-null  float64
 9   home_fouls                12680 non-null  int64  
 10  home_corners              12680 non-null  int64  
 11  home_yellowCards          12679 non-null  float64
 12  home_redCards             12680 non-null  int64  
 13  home_total_assists        12680 non-null  int64  
 14  home_total_xAssists       12680 non-null  float64
 15  home_total_key_passes     12680 non-null  int64  
 16  home_total_xGoalsChain    12680 non-null  float64
 17  home_total_xGoalsBuildup  12680 non-null  float64
 18  home_total_yellow_cards   12680 non-null  int64  
 19  home_total_red_cards      12680 non-null  int64  
 20  home_total_blocked_shots  12677 non-null  float64
 21  home_total_saved_shots    12677 non-null  float64
 22  away_xGoals               12680 non-null  float64
 23  away_shots                12680 non-null  int64  
 24  away_shotsOnTarget        12680 non-null  int64  
 25  away_deep                 12680 non-null  int64  
 26  away_ppda                 12680 non-null  float64
 27  away_fouls                12680 non-null  int64  
 28  away_corners              12680 non-null  int64  
 29  away_yellowCards          12680 non-null  float64
 30  away_redCards             12680 non-null  int64  
 31  away_total_assists        12680 non-null  int64  
 32  away_total_xAssists       12680 non-null  float64
 33  away_total_key_passes     12680 non-null  int64  
 34  away_total_xGoalsChain    12680 non-null  float64
 35  away_total_xGoalsBuildup  12680 non-null  float64
 36  away_total_yellow_cards   12680 non-null  int64  
 37  away_total_red_cards      12680 non-null  int64  
 38  away_total_blocked_shots  12672 non-null  float64
 39  away_total_saved_shots    12672 non-null  float64
dtypes: float64(16), int64(24)
memory usage: 4.0 MB
```

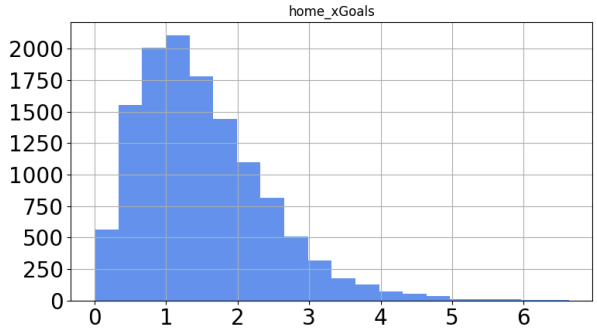
```
array([[<Axes: title={'center': 'home_Goals'}>,
        <Axes: title={'center': 'away_Goals'}>],
       [<Axes: title={'center': 'home_GoalsHalfTime'}>,
        <Axes: title={'center': 'away_GoalsHalfTime'}>],
       [<Axes: title={'center': 'home_xGoals'}>,
        <Axes: title={'center': 'home_shots'}>],
       [<Axes: title={'center': 'home_shotsOnTarget'}>,
        <Axes: title={'center': 'home_deep'}>],
       [<Axes: title={'center': 'home_ppda'}>,
        <Axes: title={'center': 'home_fouls'}>],
       [<Axes: title={'center': 'home_corners'}>,
        <Axes: title={'center': 'home_yellowCards'}>],
       [<Axes: title={'center': 'home_redCards'}>,
        <Axes: title={'center': 'home_total_assists'}>],
       [<Axes: title={'center': 'home_total_xAssists'}>,
        <Axes: title={'center': 'home_total_key_passes'}>],
       [<Axes: title={'center': 'home_total_xGoalsChain'}>,
        <Axes: title={'center': 'home_total_xGoalsBuildup'}>],
       [<Axes: title={'center': 'home_total_yellow_cards'}>,
        <Axes: title={'center': 'home_total_red_cards'}>],
       [<Axes: title={'center': 'home_total_blocked_shots'}>,
        <Axes: title={'center': 'home_total_saved_shots'}>],
       [<Axes: title={'center': 'away_xGoals'}>,
        <Axes: title={'center': 'away_shots'}>],
       [<Axes: title={'center': 'away_shotsOnTarget'}>,
        <Axes: title={'center': 'away_deep'}>],
       [<Axes: title={'center': 'away_ppda'}>,
        <Axes: title={'center': 'away_fouls'}>],
       [<Axes: title={'center': 'away_corners'}>,
        <Axes: title={'center': 'away_yellowCards'}>],
       [<Axes: title={'center': 'away_redCards'}>,
        <Axes: title={'center': 'away_total_assists'}>],
       [<Axes: title={'center': 'away_total_xAssists'}>,
        <Axes: title={'center': 'away_total_key_passes'}>],
       [<Axes: title={'center': 'away_total_xGoalsChain'}>,
        <Axes: title={'center': 'away_total_xGoalsBuildup'}>],
       [<Axes: title={'center': 'away_total_yellow_cards'}>,
        <Axes: title={'center': 'away_total_red_cards'}>],
       [<Axes: title={'center': 'away_total_blocked_shots'}>,
        <Axes: title={'center': 'away_total_saved_shots'}>]], dtype=object)
```

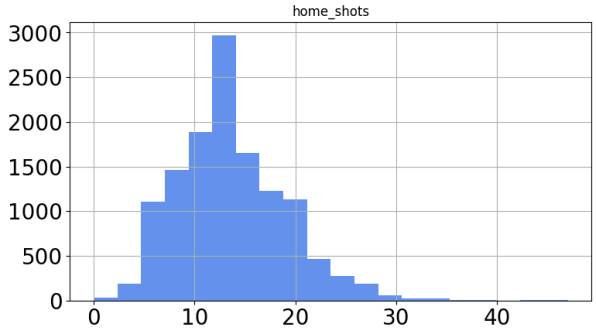

home_Goals

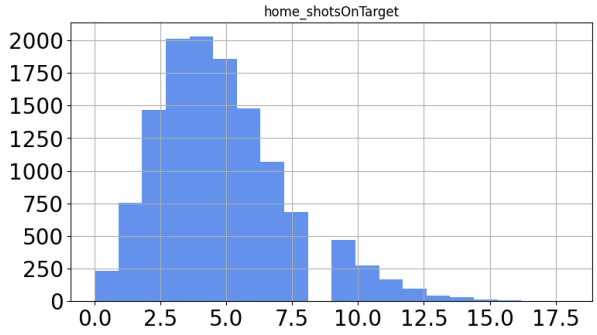

away_Goals

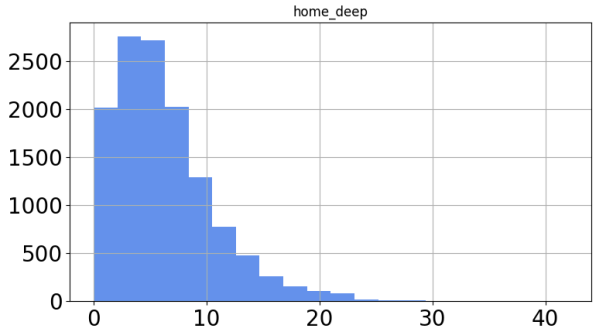

home_GoalsHalfTime

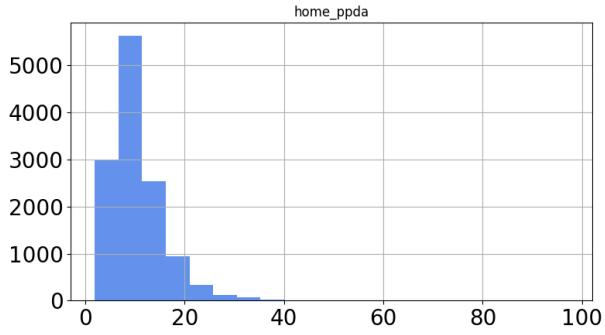

away_GoalsHalfTime

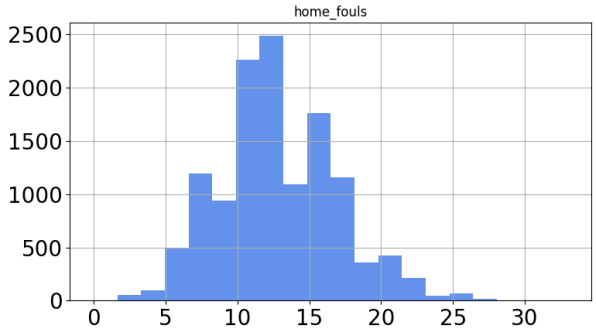

home_xGoals

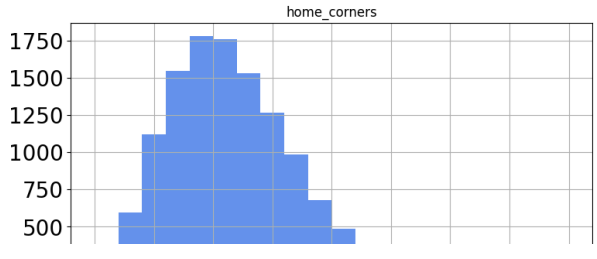

home_shots

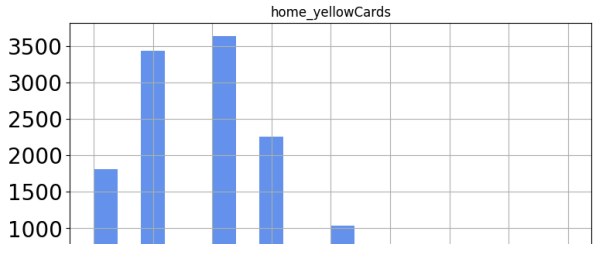

home_shotsOnTarget

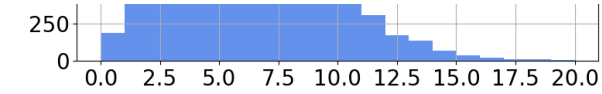

home_deep

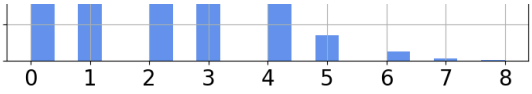

home_ppda

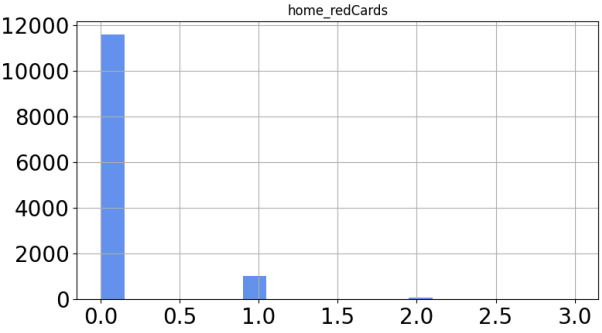

home_fouls

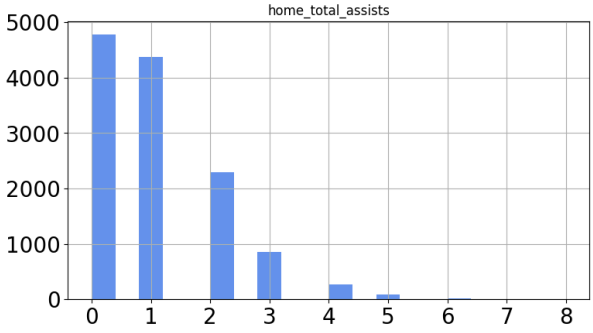

home_corners

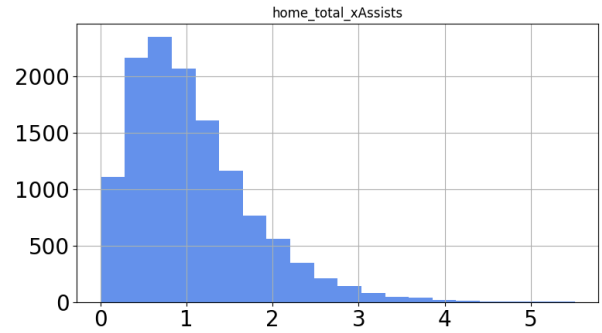

home_yellowCards

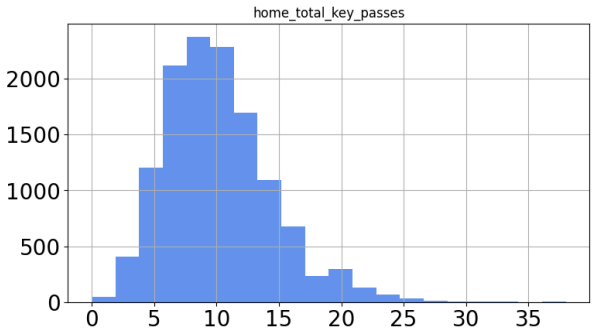

250
0
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0

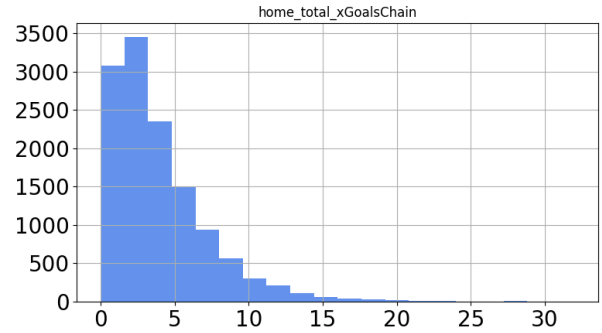

500
0
0 1 2 3 4 5 6 7 8

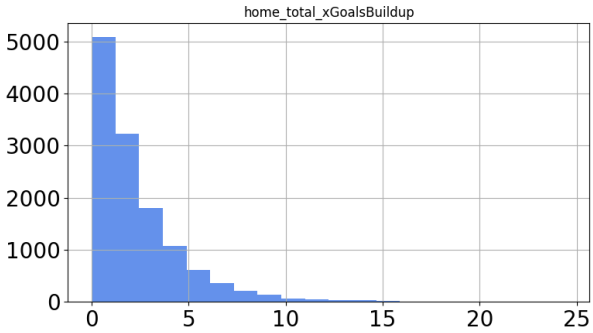

home_redCards
12000
10000
8000
6000
4000
2000
0
0.0 0.5 1.0 1.5 2.0 2.5 3.0

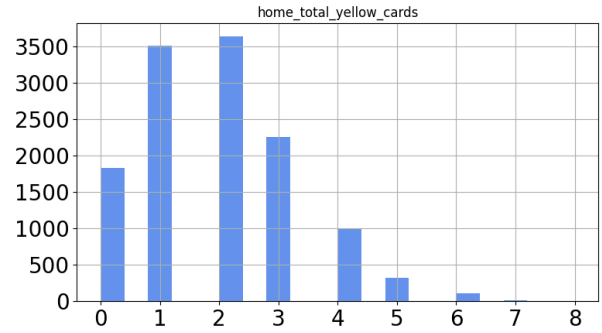

home_total_assists
5000
4000
3000
2000
1000
0
0 1 2 3 4 5 6 7 8

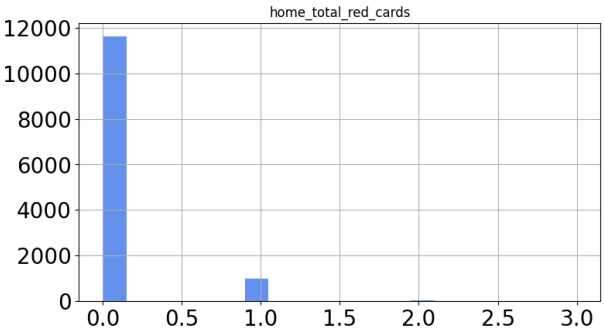

home_total_xAssists
2000
1500
1000
500
0
0 1 2 3 4 5

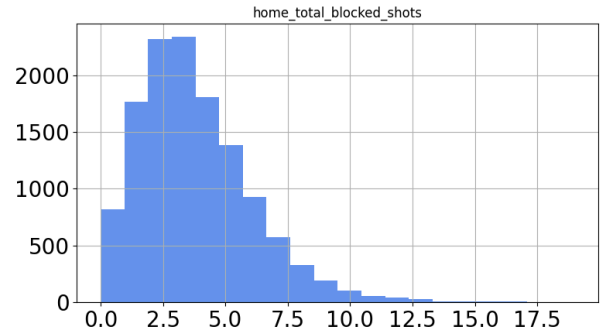

home_total_key_passes
2000
1500
1000
500
0
0 5 10 15 20 25 30 35

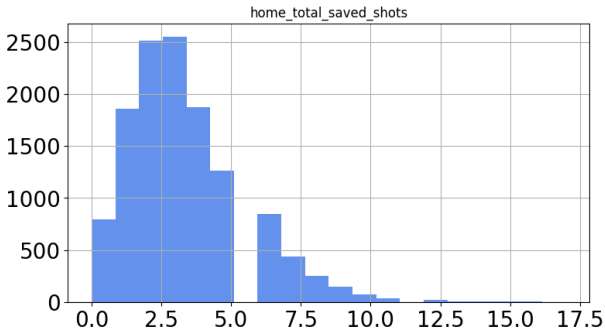

home_total_xGoalsChain
3500
3000
2500
2000
1500
1000
500
0
0 5 10 15 20 25 30


home_total_xGoalsBuildup
5000
4000
3000
2000
1000
0
0 5 10 15 20 25

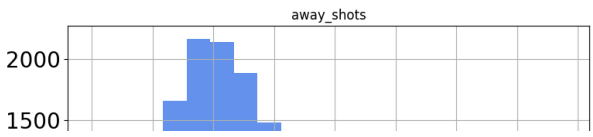

home_total_yellow_cards
3500
3000
2500
2000
1500
1000
500
0
0 1 2 3 4 5 6 7 8

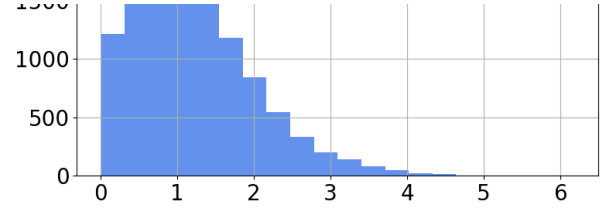

home_total_red_cards
12000
10000
8000
6000
4000
2000
0
0.0 0.5 1.0 1.5 2.0 2.5 3.0

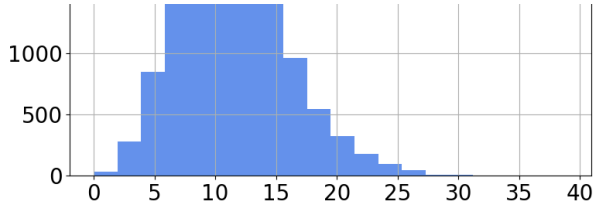

home_total_blocked_shots
2000
1500
1000
500
0
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5

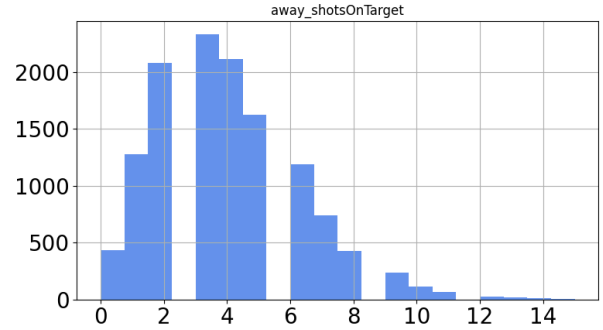

home_total_saved_shots
2500
2000
1500
1000
500
0
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5

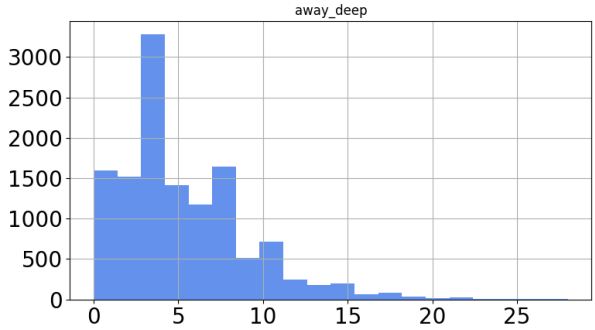

away_xGoals
2000
1500

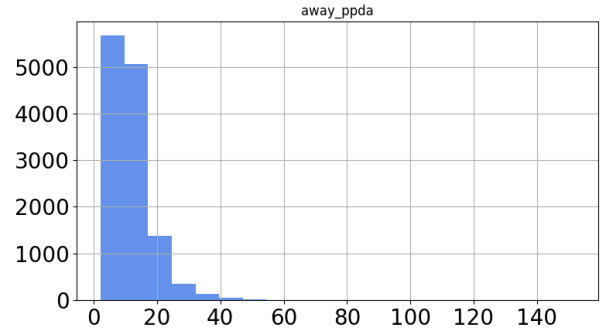

away_shots
2000
1500

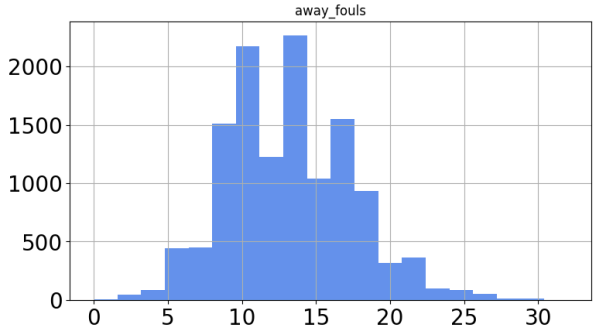


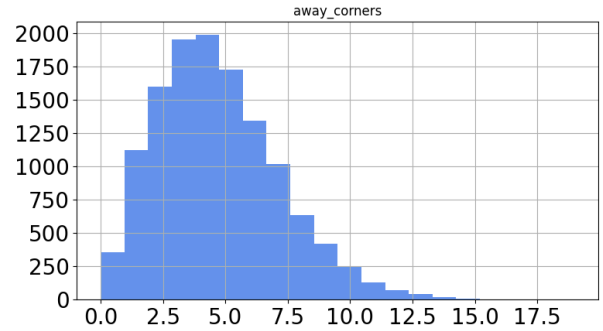


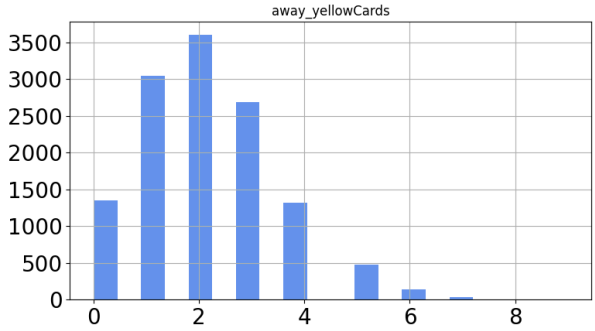

away_shotsOnTarget

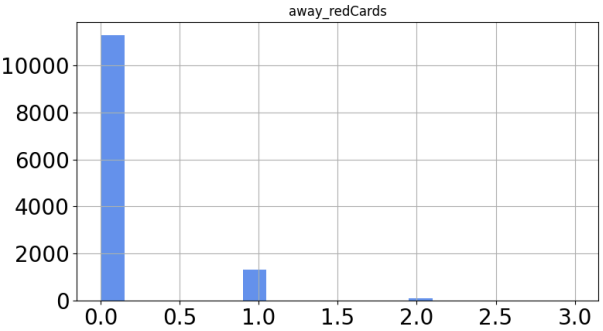

away_deep

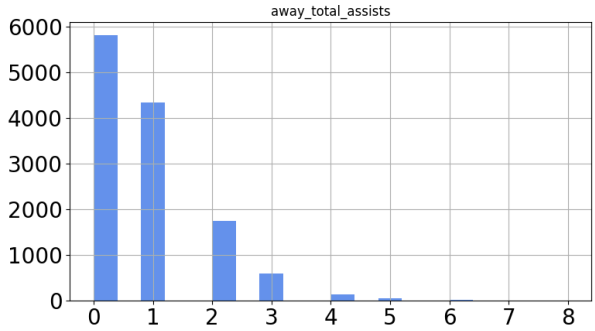

away_ppda

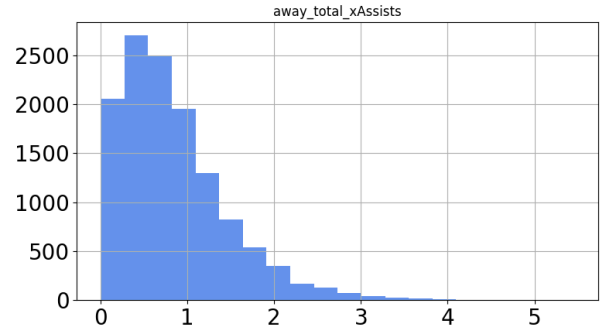

away_fouls

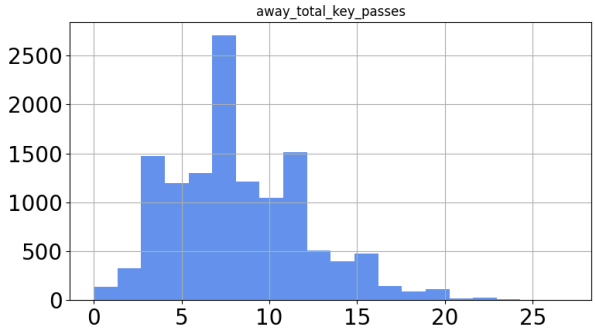

away_corners

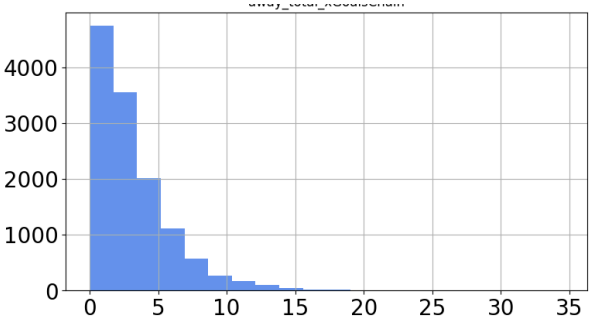

away_yellowCards

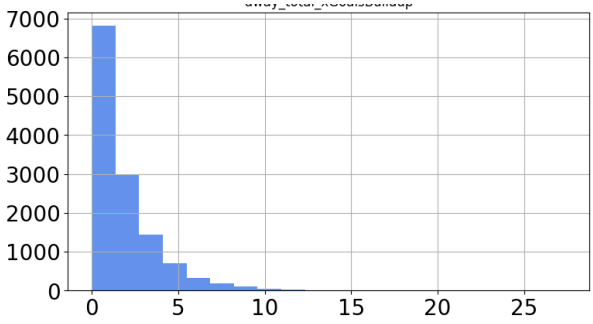

away_redCards

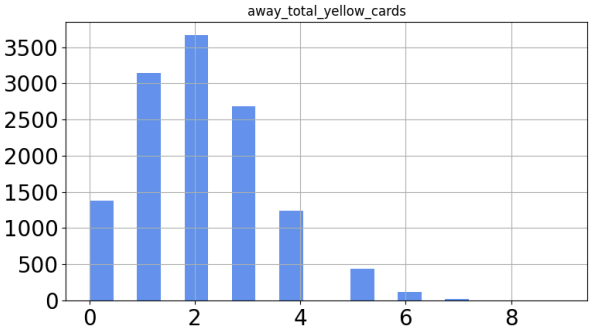

away_total_assists

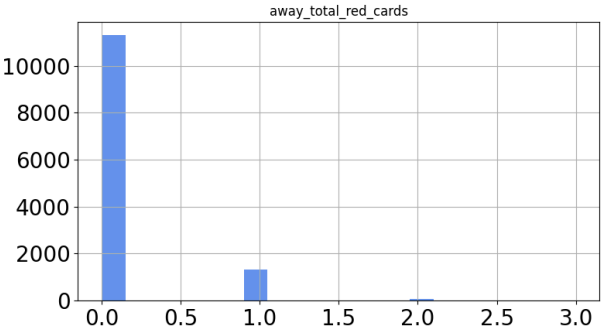

away_total_xAssists

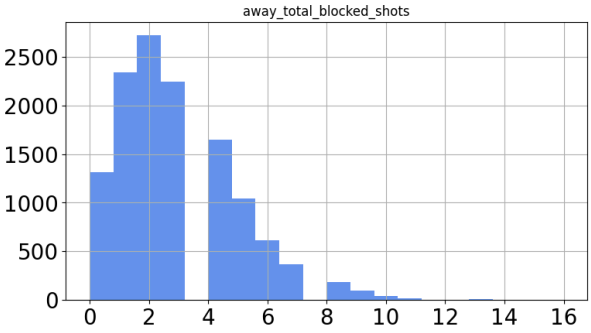

away_total_key_passes

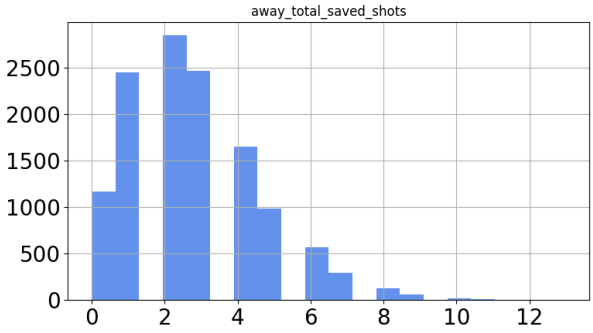





away_total_yellow_cards


away_total_red_cards


away_total_blocked_shots


away_total_saved_shots

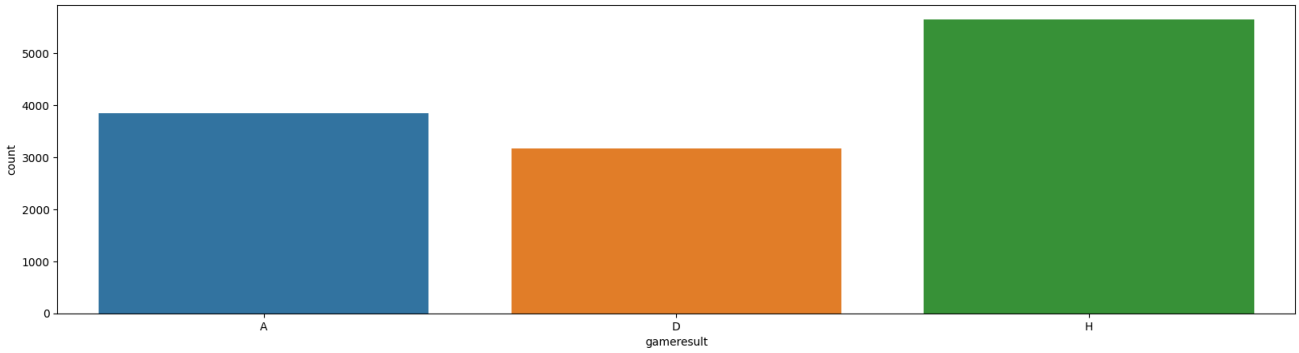
Skewness

	skewness
home_total_red_cards	3.421199
home_redCards	3.412354
away_ppda	3.277596
away_total_red_cards	2.893004
away_redCards	2.880568
home_ppda	2.602424
away_total_xGoalsBuildup	2.408524
home_total_xGoalsBuildup	2.191562
away_total_xGoalsChain	1.972202
home_total_xGoalsChain	1.819149
away_GoalsHalfTime	1.412630
away_total_xAssists	1.388455
home_deep	1.358435
away_total_assists	1.337440
home_GoalsHalfTime	1.294446
away_deep	1.282733
home_total_xAssists	1.226424
home_total_assists	1.154587
away_xGoals	1.126784
away_Goals	1.122382
home_xGoals	1.057793
away_total_blocked_shots	1.020413
home_Goals	0.990914
home_total_blocked_shots	0.988462
home_total_saved_shots	0.937809
away_total_saved_shots	0.859399
away_shotsOnTarget	0.745620
home_shotsOnTarget	0.722233
home_total_key_passes	0.701148
away_corners	0.700852
home_corners	0.698146
home_yellowCards	0.665313
away_total_key_passes	0.654491
home_shots	0.622889
home_total_yellow_cards	0.614549
away_shots	0.600380

	skewness
away_yellowCards	0.530416
away_total_yellow_cards	0.493265
home_fouls	0.387620
away_fouls	0.385974

Y - Target Value

```
<Axes: xlabel='gameresult', ylabel='count'>
```




```
H    5654
A    3854
D    3172
Name: gameresult, dtype: int64
```

Label Encoding

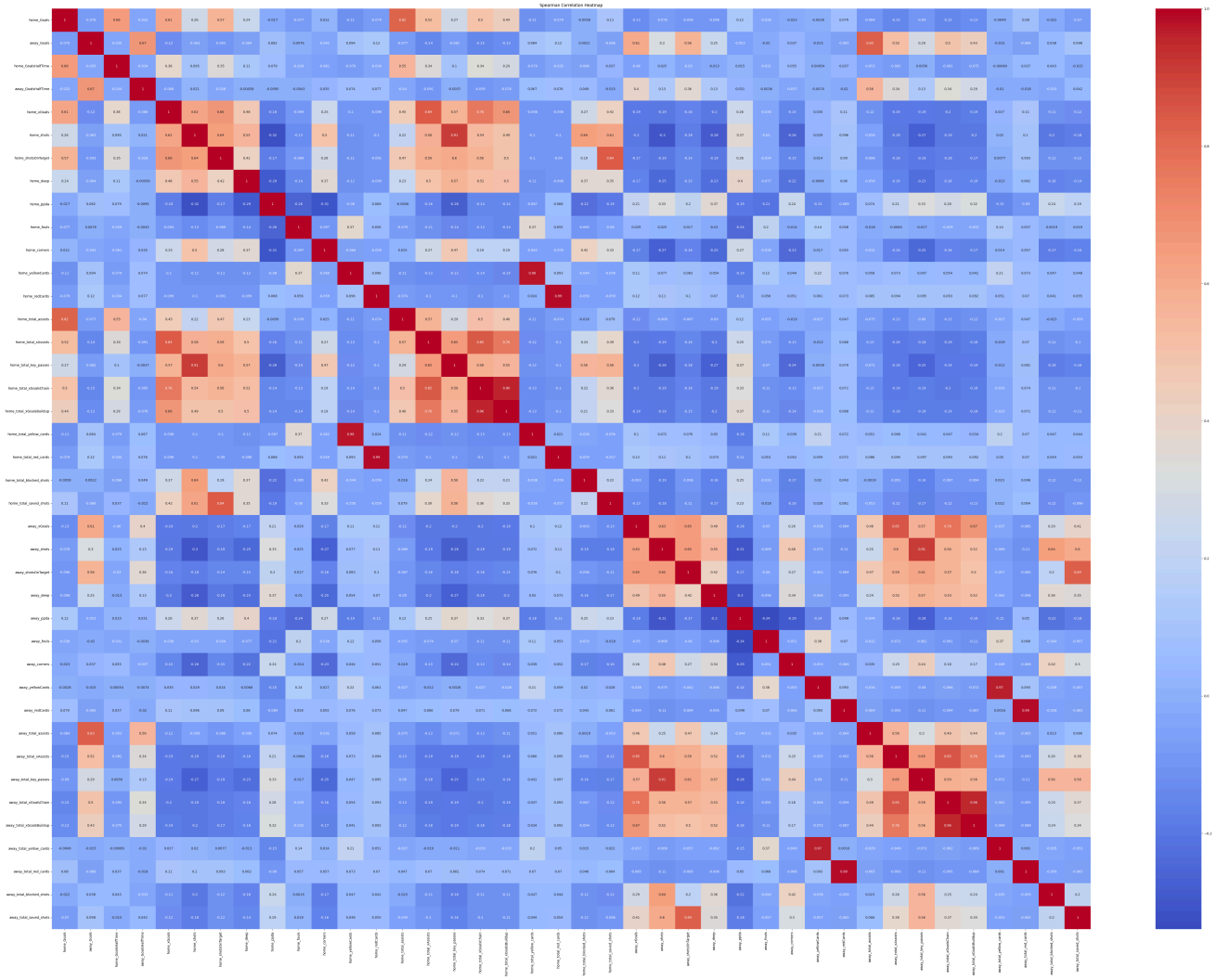
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12680 entries, 0 to 12679
Data columns (total 47 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   gameID                    12680 non-null  int64
 1   leagueID                  12680 non-null  int64
 2   season                    12680 non-null  int64
 3   date                      12680 non-null  datetime64[ns]
 4   homeTeamID                12680 non-null  int64
 5   awayTeamID                12680 non-null  int64
 6   home_Goals                12680 non-null  int64
 7   away_Goals                12680 non-null  int64
 8   home_GoalsHalfTime        12680 non-null  int64
 9   away_GoalsHalfTime        12680 non-null  int64
 10  home_xGoals               12680 non-null  float64
 11  home_shots                12680 non-null  int64
 12  home_shotsOnTarget        12680 non-null  int64
 13  home_deep                 12680 non-null  int64
 14  home_ppda                 12680 non-null  float64
 15  home_fouls                12680 non-null  int64
 16  home_corners              12680 non-null  int64
 17  home_yellowCards          12679 non-null  float64
 18  home_redCards             12680 non-null  int64
 19  home_total_assists        12680 non-null  int64
 20  home_total_xAssists       12680 non-null  float64
 21  home_total_key_passes     12680 non-null  int64
 22  home_total_xGoalsChain    12680 non-null  float64
 23  home_total_xGoalsBuildup  12680 non-null  float64
 24  home_total_yellow_cards   12680 non-null  int64
 25  home_total_red_cards      12680 non-null  int64
 26  home_total_blocked_shots  12677 non-null  float64
 27  home_total_saved_shots    12677 non-null  float64
 28  away_xGoals               12680 non-null  float64
 29  away_shots                12680 non-null  int64
 30  away_shotsOnTarget        12680 non-null  int64
 31  away_deep                 12680 non-null  int64
 32  away_ppda                 12680 non-null  float64
 33  away_fouls                12680 non-null  int64
 34  away_corners              12680 non-null  int64
 35  away_yellowCards          12680 non-null  float64
 36  away_redCards             12680 non-null  int64
 37  away_total_assists        12680 non-null  int64
 38  away_total_xAssists       12680 non-null  float64
 39  away_total_key_passes     12680 non-null  int64
 40  away_total_xGoalsChain    12680 non-null  float64
 41  away_total_xGoalsBuildup  12680 non-null  float64
 42  away_total_yellow_cards   12680 non-null  int64
 43  away_total_red_cards      12680 non-null  int64
 44  away_total_blocked_shots  12672 non-null  float64
 45  away_total_saved_shots    12672 non-null  float64
 46  gameresult                12680 non-null  int64
dtypes: datetime64[ns](1), float64(16), int64(30)
memory usage: 4.6 MB
```

Correlation

	home_Goals	away_Goals	home_GoalsHalfTime	away_GoalsHalfTime	h
home_Goals	1.000000	-0.075874	0.658365	-0.031795	
away_Goals	-0.075874	1.000000	-0.034998	0.665542	
home_GoalsHalfTime	0.658365	-0.034998	1.000000	-0.034418	
away_GoalsHalfTime	-0.031795	0.665542	-0.034418	1.000000	
home_xGoals	0.608347	-0.124532	0.384339	-0.067653	
home_shots	0.264244	-0.062635	0.094571	0.020847	
home_shotsOnTarget	0.566983	-0.093423	0.347701	-0.028309	
home_deep	0.244392	-0.064400	0.113474	-0.000579	
home_ppda	-0.016869	0.081781	0.078865	-0.009549	
home_fouls	-0.076529	0.007621	-0.028519	-0.004332	
home_corners	0.012084	-0.042092	-0.061367	0.034687	
home_yellowCards	-0.110358	0.094092	-0.078678	0.074368	
home_redCards	-0.078921	0.119724	-0.034297	0.076753	
home_total_assists	0.821408	-0.076967	0.553991	-0.040397	
home_total_xAssists	0.524452	-0.144809	0.342967	-0.090912	
home_total_key_passes	0.265327	-0.082397	0.104662	-0.003695	
home_total_xGoalsChain	0.502133	-0.148832	0.335959	-0.094924	
home_total_xGoalsBuildup	0.436624	-0.130675	0.291998	-0.076459	
home_total_yellow_cards	-0.106708	0.084468	-0.078712	0.066896	
home_total_red_cards	-0.079071	0.119791	-0.034715	0.075612	
home_total_blocked_shots	-0.005759	0.002203	-0.065756	0.048534	
home_total_saved_shots	0.108625	-0.066179	0.037200	-0.014952	
away_xGoals	-0.133065	0.606813	-0.060373	0.402162	
away_shots	-0.075582	0.301792	0.024601	0.133999	
away_shotsOnTarget	-0.095580	0.563273	-0.029802	0.362111	
away_deep	-0.098196	0.250748	-0.012637	0.131121	
away_ppda	0.124819	-0.053056	0.015049	0.030521	
away_fouls	-0.038290	-0.020160	-0.031166	-0.003575	
away_corners	-0.022524	0.036902	0.055099	-0.037495	
away_yellowCards	-0.002762	-0.015331	0.000536	-0.007396	
away_redCards	0.078895	-0.065490	0.036997	-0.019710	
away_total_assists	-0.083797	0.826960	-0.052697	0.556684	
away_total_xAssists	-0.148038	0.519682	-0.081793	0.342887	
away_total_key_passes	-0.090315	0.288618	0.005640	0.131941	
away_total_xGoalsChain	-0.153945	0.497257	-0.091736	0.336698	
away_total_xGoalsBuildup	-0.134761	0.428352	-0.075297	0.286618	

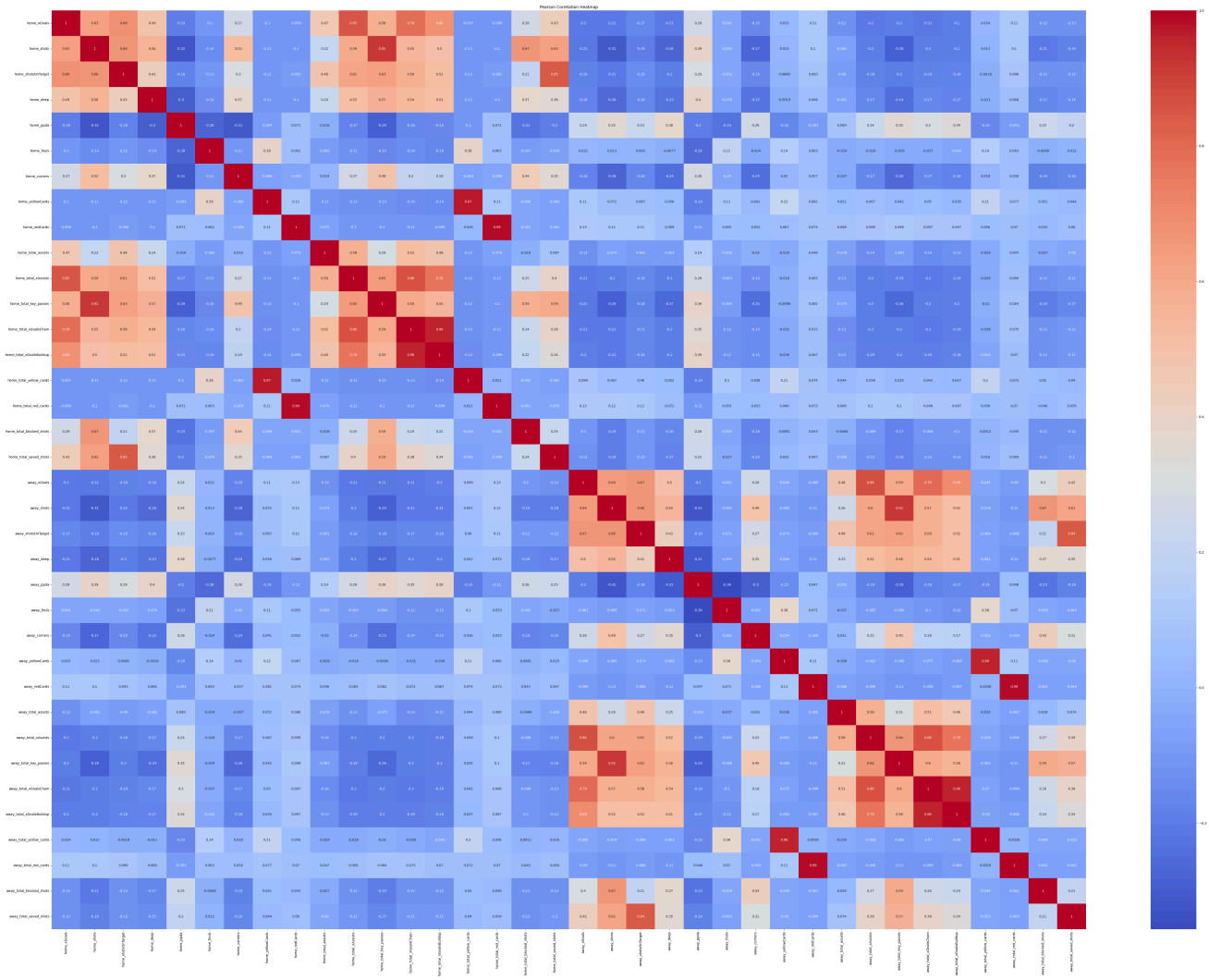
	home_Goals	away_Goals	home_GoalsHalfTime	away_GoalsHalfTime	h
away_total_yellow_cards	-0.004914	-0.015292	-0.000887	-0.009993	
away_total_red_cards	0.080064	-0.065757	0.037384	-0.017594	
away_total_blocked_shots	-0.022438	0.038495	0.042827	-0.033448	
away_total_saved_shots	-0.069618	0.098243	-0.014708	0.041945	

40 rows × 40 columns




```
Skewed columns: ['home_xGoals', 'home_deep', 'home_ppda', 'home_redCards', 'home_t
otal_assists', 'home_total_xAssists', 'home_total_xGoalsChain', 'home_total_xGoals
Buildup', 'home_total_red_cards', 'away_xGoals', 'away_deep', 'away_ppda', 'away_r
edCards', 'away_total_assists', 'away_total_xAssists', 'away_total_xGoalsChain',
'away_total_xGoalsBuildup', 'away_total_red_cards', 'away_total_blocked_shots']
```

	home_xGoals	home_shots	home_shotsOnTarget	home_deep	home_pp
home_xGoals	1.000000	0.629700	0.676205	0.490572	-0.190
home_shots	0.629700	1.000000	0.664302	0.555342	-0.323
home_shotsOnTarget	0.676205	0.664302	1.000000	0.428827	-0.178
home_deep	0.490572	0.555342	0.428827	1.000000	-0.296
home_ppda	-0.190396	-0.323308	-0.178512	-0.296802	1.000
home_fouls	-0.100012	-0.143358	-0.107106	-0.136951	-0.281
home_corners	0.266592	0.521157	0.298436	0.370490	-0.313
home_yellowCards	-0.100178	-0.111155	-0.115629	-0.111548	-0.094
home_redCards	-0.098994	-0.100345	-0.091992	-0.104962	0.071
home_total_assists	0.465161	0.217487	0.477356	0.242030	-0.016
home_total_xAssists	0.854483	0.587158	0.608535	0.517134	-0.171
home_total_key_passes	0.583735	0.919711	0.627367	0.569922	-0.287
home_total_xGoalsChain	0.782750	0.549993	0.579888	0.535408	-0.160
home_total_xGoalsBuildup	0.683526	0.498679	0.518232	0.511825	-0.150
home_total_yellow_cards	-0.097441	-0.105691	-0.110396	-0.105468	-0.101
home_total_red_cards	-0.098842	-0.099813	-0.090929	-0.104991	0.071
home_total_blocked_shots	0.279103	0.668401	0.206916	0.371457	-0.228
home_total_saved_shots	0.434595	0.631525	0.846096	0.359634	-0.196
away_xGoals	-0.203855	-0.209873	-0.184316	-0.175495	0.236
away_shots	-0.205391	-0.309636	-0.205030	-0.259221	0.351
away_shotsOnTarget	-0.167961	-0.189016	-0.149913	-0.160664	0.226
away_deep	-0.206215	-0.279483	-0.199017	-0.232322	0.375
away_ppda	0.278413	0.385874	0.294317	0.399520	-0.195
away_fouls	-0.040590	-0.042263	-0.052931	-0.075542	-0.230
away_corners	-0.151564	-0.268496	-0.153183	-0.221489	0.259
away_yellowCards	0.033374	0.023254	0.006633	-0.001940	-0.156
away_redCards	0.108958	0.100343	0.093368	0.065757	-0.092
away_total_assists	-0.120000	-0.060860	-0.090249	-0.062478	0.088
away_total_xAssists	-0.203569	-0.199048	-0.182726	-0.169032	0.235
away_total_key_passes	-0.201094	-0.282818	-0.198427	-0.240961	0.350
away_total_xGoalsChain	-0.208703	-0.201666	-0.186717	-0.170316	0.304
away_total_xGoalsBuildup	-0.198090	-0.200045	-0.178356	-0.174496	0.337
away_total_yellow_cards	0.024007	0.012422	-0.001791	-0.010788	-0.154
away_total_red_cards	0.112732	0.103960	0.095703	0.067784	-0.093
away_total_blocked_shots	-0.121896	-0.208027	-0.125841	-0.168472	0.252
away_total_saved_shots	-0.126465	-0.186125	-0.120483	-0.145080	0.204



1. Why Two Correlation Matrices?

Spearman's Correlation:

Measures monotonic relationships (if one variable goes up, does the other consistently go up or down?). Rank-based, so it's more robust to outliers and skew. A high Spearman correlation (ϱ ρ close to +1 or –1) means the two variables move together in rank, even if the relationship isn't strictly linear. Pearson's Correlation:

Measures linear relationships. More sensitive to outliers and skewed data. A high Pearson correlation (r r close to +1 or –1) means the two variables move together in a roughly linear fashion. By comparing them, you see where variables are consistently related (both Spearman and Pearson are large in magnitude) versus where the relationship might be non-linear (Spearman is strong, but Pearson is weaker) or influenced by outliers/skew.

2. Identify Strong Relationships

Look for correlation coefficients with large absolute values (e.g., above ~0.5 or 0.6). For instance, in your Spearman table, you might see:

home_xGoals is strongly correlated (Spearman ≈ 0.84 ≈0.84) with home_total_xAssists, suggesting that as the home team's expected goals increase, so do their total xAssists in a fairly monotonic way. Similarly, in Pearson's matrix, home_xGoals and home_total_xAssists also show a strong linear correlation (~0.85). When a relationship is consistently strong in both Spearman and Pearson, that often means you have a robust, near-linear association.

3. Spot Differences Due to Skew

Some variables might show a higher correlation under Spearman than Pearson (or vice versa). This often indicates:

Skew or Outliers are influencing Pearson, making the linear correlation weaker or stronger than the monotonic trend. For instance, you might see that home_ppda correlates differently with certain offensive stats in Spearman vs. Pearson. If the difference is large, it can mean a few extreme values are affecting the linear measure.

4. Multicollinearity Concerns

If you see multiple columns with correlations near ±0.8 or ±0.9, that suggests multicollinearity. For example, home_shots and home_total_key_passes might be strongly correlated. In modeling:

Dropping one of the highly correlated features or combining them (e.g., via PCA) can help reduce redundancy. Check both Spearman and Pearson for strong clusters of correlated features.

5. Modeling Implications

Feature Selection: If two variables are almost duplicates (like home_total_xGoalsChain vs. home_total_xGoalsBuildup with correlation ~0.95), you might choose only one to avoid redundancy. Transformations: If you see big differences between Spearman and Pearson, consider log transforms or other transformations on heavily skewed variables (like home_ppda or home_total_red_cards) before using them in a linear model. Interpretation: Spearman is telling you which pairs move together in rank order—good for capturing monotonic trends. Pearson is telling you which pairs have a more direct linear relationship—useful for linear regression assumptions.

6. Next Steps

Highlight Strong Correlations: Identify pairs with |corr| ≥ 0.7 or 0.8. Investigate if they're truly redundant or if each has unique predictive value. Check for Potential Data Issues: Extremely high correlation can signal duplicates or derived columns (like home_shots vs. home_total_blocked_shots might be partly overlapping). Consider Log Transform: For columns with high skew (e.g., red cards, which are often 0 or 1 with

occasional higher values), a log transform or another approach (like a Box-Cox transform) might help linear models. Re-check Pearson correlation after transformation if linear relationships are of interest. Use Spearman for Rank-Based Insights: If your target variable (gameresult) is ordinal or you suspect non-linear but monotonic relationships, Spearman can be more informative.

Summary

You have two correlation matrices—Spearman and Pearson—because your data is skewed, and you want to compare monotonic vs. linear relationships. High correlation in both indicates a strong linear relationship that's robust to skew. Differences between Spearman and Pearson can highlight skew/outliers or non-linear patterns. Use these insights to choose features, consider transformations, and interpret whether variables have linear or simply monotonic relationships. Ultimately, both matrices are valuable: Spearman for robust, monotonic insight and Pearson for linear modeling considerations.

T-Test

```
T-statistic: 51.8705, p-value: 0.0000
```

You can confidently reject the null hypothesis that there is no difference in home_xGoals between home wins and non-home wins. In practical terms, this means that the home team's expected goals (home_xGoals) are significantly higher (or lower, depending on the direction of the difference) when they win compared to when they don't win, and this difference is not due to random chance. If you're using this variable for predictive modeling or further analysis, it appears to be a very strong indicator of match outcome in your dataset.

```
ANOVA F-statistic: 1547.4971, p-value: 0.0000
```

This indicates that the means of the home_xGoals variable differ significantly across all three groups (home win, draw, away win). In other words, at least one group's mean is different from the others. This justifies that home_xGoals is an important metric in differentiating match outcomes.

```
Paired t-test: t-statistic = 28.1871, p-value = 0.0000
```

The paired t-test shows a highly significant difference between home_xGoals and away_xGoals for each game. This suggests that the disparity between home and away expected goals is very informative about the match outcome.

```
ANOVA for home_shots: F-statistic = 220.7895, p-value = 0.0000
ANOVA for away_shots: F-statistic = 331.2351, p-value = 0.0000
ANOVA for home_deep: F-statistic = 224.1493, p-value = 0.0000
ANOVA for away_deep: F-statistic = 333.4027, p-value = 0.0000
ANOVA for home_ppda: F-statistic = 55.9803, p-value = 0.0000
ANOVA for away_ppda: F-statistic = 81.4031, p-value = 0.0000
ANOVA for home_fouls: F-statistic = 22.5709, p-value = 0.0000
ANOVA for away_fouls: F-statistic = 29.2506, p-value = 0.0000
ANOVA for home_corners: F-statistic = 11.3146, p-value = 0.0000
ANOVA for away_corners: F-statistic = 4.2018, p-value = 0.0150
ANOVA for home_yellowCards: F-statistic = 84.8338, p-value = 0.0000
ANOVA for away_yellowCards: F-statistic = 16.2915, p-value = 0.0000
ANOVA for home_redCards: F-statistic = 108.9195, p-value = 0.0000
ANOVA for away_redCards: F-statistic = 74.0203, p-value = 0.0000
```

These ANOVA results tell you that for every feature you tested, the mean value differs significantly across the match outcome groups (gameresult), meaning that the variation between groups is far greater than the variation within each group. In other words:

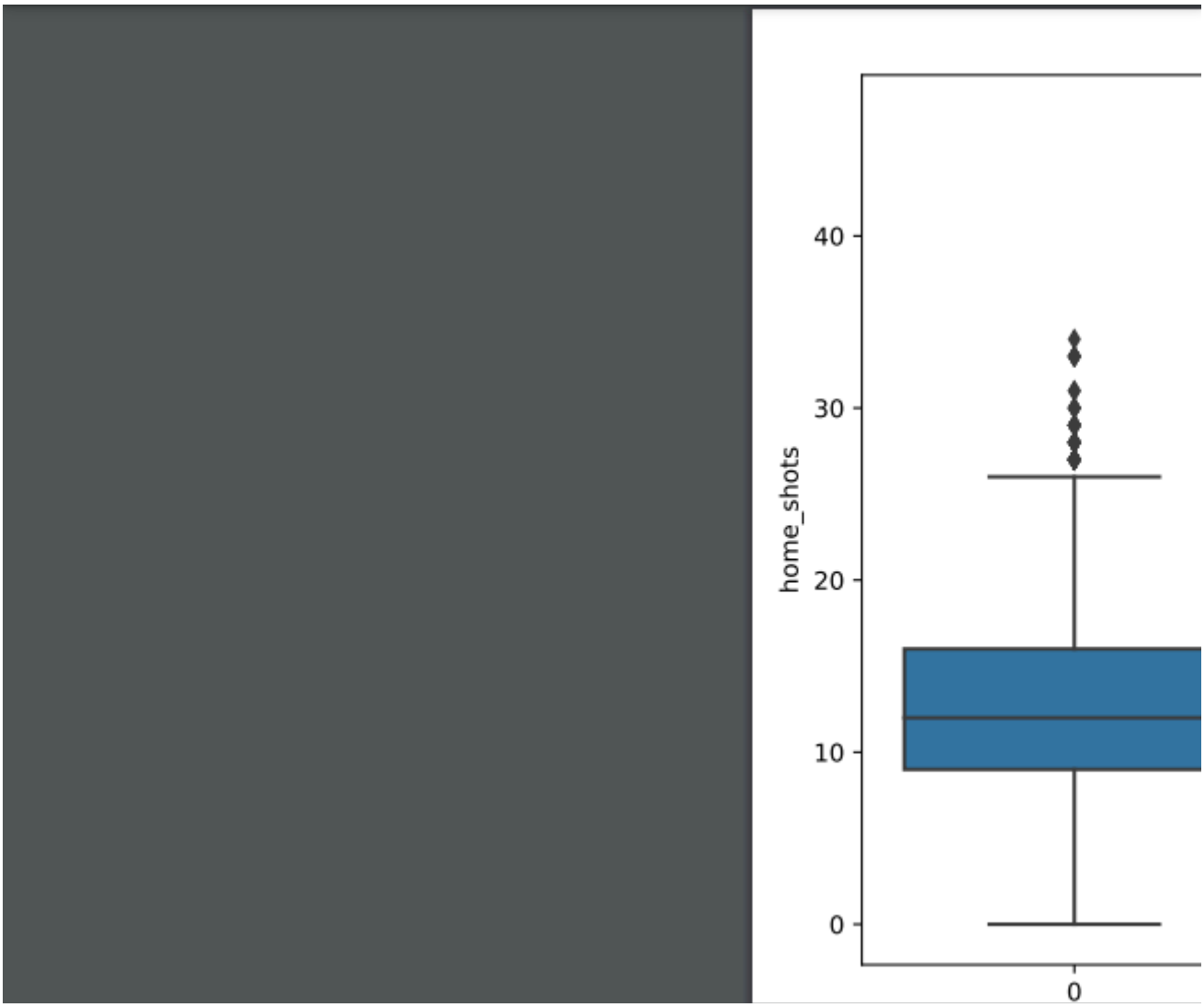
High F-statistics & Near-Zero P-values: Almost every feature (e.g., home_shots, away_shots, homeGoalsHalfTime, etc.) shows a very high F-statistic and a p-value effectively equal to 0. This indicates that the probability of observing such differences by chance is extremely low. For instance, homeGoalsHalfTime and awayGoalsHalfTime have F-statistics over 1400, suggesting that these variables differ drastically across the different match outcomes.

Away_corners Exception: Although the F-statistic for away_corners is lower (4.2018), its p-value is still below a typical significance threshold (p = 0.0150), meaning that even this feature shows statistically significant differences between groups.

```
2    5654
0    3854
1    3172
Name: gameresult, dtype: int64
```

0 for away win, 1 for draw, 2 for home win

analysis_results_table.pdf



Overall Observations Most Features Show Statistically Significant Differences For nearly all features, Tukey's HSD indicates at least some pairwise group differences with $p < 0.05$, meaning the mean values of these features differ across the three gameresult categories.

Variance Differences in Many Cases Levene's test often returns a very small p-value (e.g., 0.0000), suggesting the assumption of equal variances (homoscedasticity) does not hold for a number of these features. A few features (like home_fouls, away_fouls) do not exhibit significant variance differences.

Some Features Differ Between All Pairs; Others Only Some

All pairs differ: home_shots, away_shots, home_deep, away_deep, home_redCards, away_redCards, homeGoalsHalfTime, and awayGoalsHalfTime each show significant mean differences among all three pairs (0 vs. 1, 0 vs. 2, and 1 vs. 2). Partial differences: Some features only differ for certain group pairs. For example, home_ppda: differs between (0 vs. 1) and (0 vs. 2), but not (1 vs. 2). home_fouls: differs between (0 vs. 2) and (1 vs. 2), but not (0 vs. 1). away_fouls: differs between (0 vs. 1) and (1 vs. 2), but not

(0 vs. 2). Interpreting “Reject=True” Whenever Tukey’s HSD shows “reject = True,” it means there is a statistically significant difference in the mean of that feature between those two gameresult groups at the chosen alpha level (0.05 by default).

High-Level Takeaways Shots (home_shots, away_shots) and Deep (home_deep, away_deep) are strongly different among all outcomes, indicating that teams’ shot counts and attacking penetration vary considerably depending on the result. Half-Time Goals (homeGoalsHalfTime, awayGoalsHalfTime) also differ significantly for all pairs, suggesting that scoring patterns before halftime strongly correlate with the final match outcome. Red Cards (home_redCards, away_redCards) differ across all three groups, but the differences are quite small in absolute terms. Still, they are statistically meaningful. PPDA (home_ppda, away_ppda) shows partial differences. For instance, away_ppda is significantly different between (0 vs. 2) and (1 vs. 2) but not between (0 vs. 1). Fouls and Corners often differ for two of the pairs but not all three, indicating these stats may be somewhat less discriminative than others like shots or goals. Variance Differences: In many cases (e.g., home_ppda, away_ppda, homeShots), Levene’s test reveals that the spread of values is not the same across all outcome groups, meaning you should be cautious if you assume homoscedasticity for further parametric tests. Practical Implications Predictive Modeling: Features such as shots, xGoals (if available), deep completions, and first-half goals appear highly indicative of final match outcomes. They may serve as strong predictors in classification models.

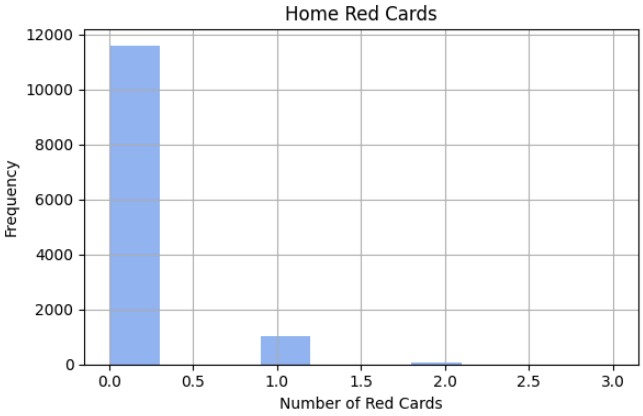
Further Analysis:

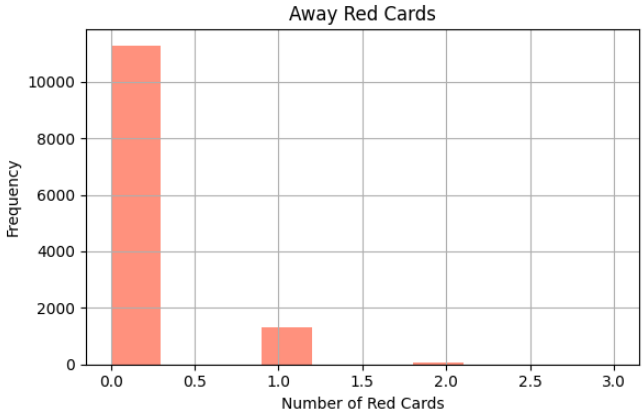
Pairwise differences from Tukey’s HSD highlight where exactly the mean of a feature is higher or lower. For example, if “group2” (possibly home wins) consistently has higher home_shots than “group0” or “group1,” that indicates the home team’s shot volume is a good indicator of a home win. Variance differences from Levene’s test mean you should check assumptions (e.g., for ANOVA or regression). If variances are unequal, you might use Welch’s ANOVA or heteroscedastic-robust methods. Game Strategy Insights:

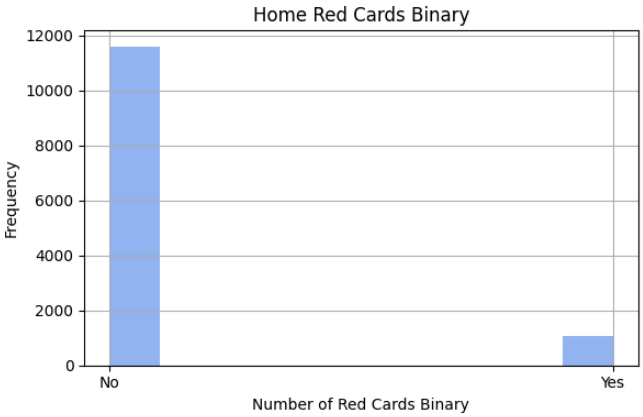
High shot counts and high attacking penetration (deep completions) are strongly tied to winning or losing. Early goals (reflected in half-time goals) also have a major effect on eventual outcomes. Discipline stats (cards, fouls) do show some differences but are generally not as consistently discriminative across all pairs as shots and goals are. Conclusion The PDF results confirm that most match statistics differ significantly across the three outcome categories. While some features (e.g., home_ppda, home_fouls) only differ for specific pairwise comparisons, many others (shots, deep completions, half-time goals) show universal differences among all three groups. Furthermore, unequal variances are common, so any parametric modeling approach should account for heteroscedasticity or use robust methods.

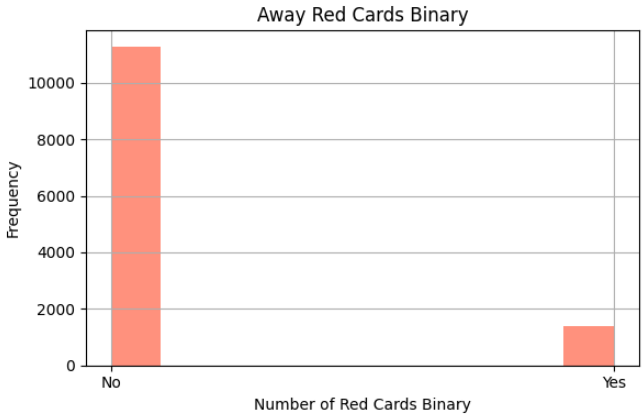
Overall, these findings suggest that the identified features—particularly shots, half-time goals, and deep completions—are strongly associated with match outcomes and could be valuable for predictive or explanatory models of football results.

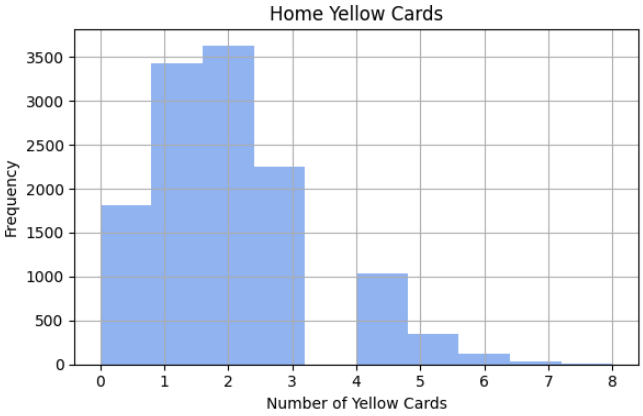
Chi-Square

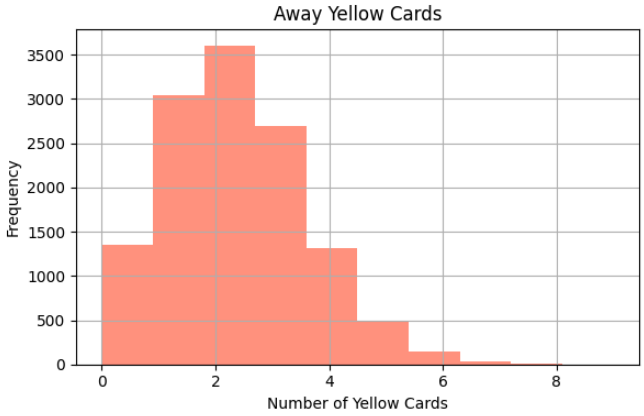












	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
4140	4888	2	2014	2015-03-02 19:45:00	95	98	1	1

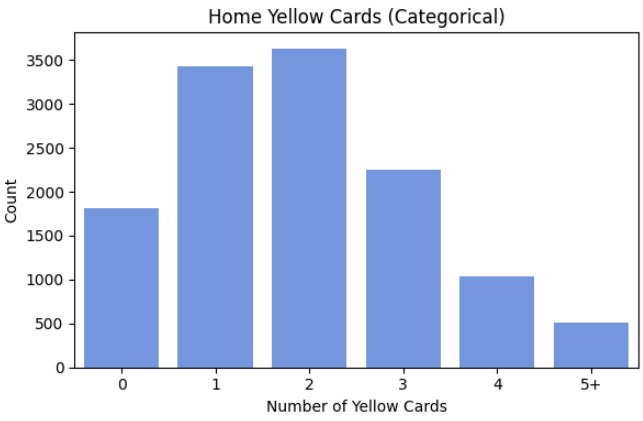
1 rows × 49 columns

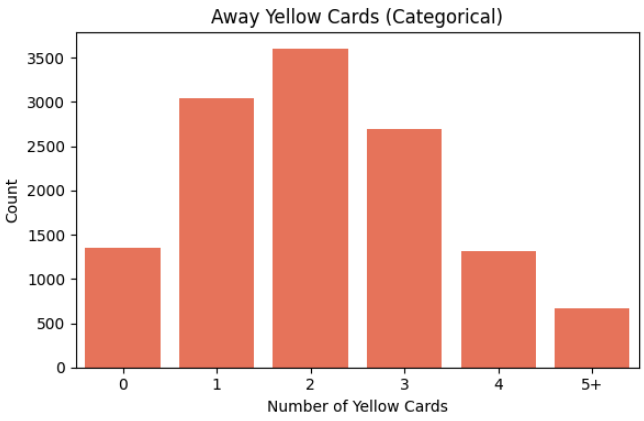
gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals	home_G

0 rows × 49 columns

```
2.0    3630
1.0    3433
3.0    2252
0.0    1813
4.0    1038
5.0     348
6.0     128
7.0      30
8.0       7
Name: home_yellowCards, dtype: int64
2.0    3605
1.0    3045
3.0    2691
0.0    1356
4.0    1319
5.0     480
6.0     141
7.0      35
8.0       6
9.0       2
Name: away_yellowCards, dtype: int64


2     3631
1     3433
3     2252
0     1813
4     1038
5+     513
Name: home_yellowCards_cat, dtype: int64
2     3605
1     3045
3     2691
0     1356
4     1319
5+     664
Name: away_yellowCards_cat, dtype: int64
```





```
=== Chi-square test: gameresult vs. leagueID ===
Contingency Table:
leagueID       1     2    3     4     5
gameresult
0            845   837  651   764   757
1            629   668  532   682   661
2           1186  1155  959  1214  1140

Expected Frequencies:
leagueID              1            2           3            4            5
gameresult
0            808.488959   808.488959  651.046372   808.488959   777.486751
1            665.419558   665.419558  535.837855   665.419558   639.903470
2           1186.091483  1186.091483  955.115773  1186.091483  1140.609779

Chi2 Statistic: 10.2695
p-value: 0.2466
Degrees of Freedom: 8
--------------------------------------------------

=== Chi-square test: gameresult vs. home_yellowCards_cat ===
Contingency Table:
home_yellowCards_cat     0     1     2    3    4   5+
gameresult
0                      438   977  1161  731  361  186
1                      372   775   949  622  320  134
2                     1003  1681  1521  899  357  193

Expected Frequencies:
home_yellowCards_cat           0            1            2            3  \
gameresult
0                     551.049054  1043.437066  1103.617823   684.480126
1                     453.535962   858.791483   908.322713   563.355205
2                     808.414984  1530.771451  1619.059464  1004.164669

home_yellowCards_cat           4          5+
gameresult
0                     315.493060  155.922871
1                     259.663722  128.330915
2                     462.843218  228.746215

Chi2 Statistic: 199.2871
p-value: 0.0000
Degrees of Freedom: 10
--------------------------------------------------

=== Chi-square test: gameresult vs. away_yellowCards_cat ===
Contingency Table:
away_yellowCards_cat    0     1     2     3    4   5+
gameresult
0                     472   969  1050   798  359  206
1                     290   707   910   688  392  185
2                     594  1369  1645  1205  568  273

Expected Frequencies:
away_yellowCards_cat           0            1            2            3  \
gameresult
0                     412.147003   925.507098  1095.715300   817.911199
1                     339.213880   761.730284   901.818612   673.174448
```

```
2                    604.639117  1357.762618  1607.466088  1199.914353

away_yellowCards_cat           4          5+
gameresult
0                    400.901104  201.818297
1                    329.958044  166.104732
2                    588.140852  296.076972

Chi2 Statistic: 46.5489
p-value: 0.0000
Degrees of Freedom: 10
--------------------------------------------------

=== Chi-square test: gameresult vs. home_redCards_binary ===
Contingency Table:
home_redCards_binary    No  Yes
gameresult
0                    3345  509
1                    2878  294
2                    5379  275

Expected Frequencies:
home_redCards_binary           No         Yes
gameresult
0                    3526.349211  327.650789
1                    2902.329968  269.670032
2                    5173.320820  480.679180

Chi2 Statistic: 208.2850
p-value: 0.0000
Degrees of Freedom: 2
--------------------------------------------------

=== Chi-square test: gameresult vs. away_redCards_binary ===
Contingency Table:
away_redCards_binary    No  Yes
gameresult
0                    3605  249
1                    2828  344
2                    4851  803

Expected Frequencies:
away_redCards_binary           No         Yes
gameresult
0                    3429.695268  424.304732
1                    2822.779811  349.220189
2                    5031.524921  622.475079

Chi2 Statistic: 140.3080
p-value: 0.0000
Degrees of Freedom: 2
--------------------------------------------------
```

1. gameresult vs. leagueID

p-value = 0.2466 (above 0.05) There is no statistically significant association between the match outcome (gameresult) and leagueID. In other words, the distribution of

home wins/draws/away wins doesn’t differ enough across leagues to be considered non-random at the 5% significance level.

2. gameresult vs. home_yellowCards_cat

p-value = 0.0000 (well below 0.05) Statistically significant association exists between game result and the binned home yellow-card categories (0, 1, 2, 3, 4, 5+). This implies that the distribution of home yellow cards (how many the home team gets) is not independent of whether the home team ended up winning, drawing, or losing.

3. gameresult vs. away_yellowCards_cat

p-value = 0.0000 Similarly, a significant association between game result and the binned away yellow-card categories. The number of yellow cards the away team receives is not independent of the match outcome.

4. gameresult vs. home_redCards_binary

p-value = 0.0000 There’s a significant association between game result and whether the home team had no red cards (No) or at least one red card (Yes). In simpler terms, whether the home team sees red cards correlates with whether the home team won, drew, or lost.

5. gameresult vs. away_redCards_binary

p-value = 0.0000 Another significant association: the away team’s red-card status (No vs. Yes) is tied to the final outcome.

Overall Takeaways

LeagueID does not appear to affect the distribution of match outcomes (H/D/A). Cards (both yellow and red, for home and away) do show a statistically significant association with match outcomes. This does not tell you which categories lead to more wins or losses, only that they are not independent. For directionality or deeper insight, you could: Look at residuals (which categories are over- or under-represented). Conduct post-hoc tests (e.g., pairwise comparisons). Examine the contingency tables in more detail (e.g., more red cards in losing teams?). Because p-values for all card-related tests are effectively zero, you can conclude that the distribution of match outcomes is strongly associated with the number of cards (yellow or red) teams receive.

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
0	81	1	2015	2015-08-08 15:45:00	89	82	1	0
1	82	1	2015	2015-08-08 18:00:00	73	71	0	1
2	83	1	2015	2015-08-08 18:00:00	72	90	2	2
3	84	1	2015	2015-08-08 18:00:00	75	77	4	2
4	85	1	2015	2015-08-08 18:00:00	79	78	1	3
...	...	...	...	...	...	...	...	...
12675	16131	5	2020	2021-05-23 19:00:00	168	166	1	2
12676	16132	5	2020	2021-05-23 19:00:00	177	176	1	2
12677	16133	5	2020	2021-05-23 19:00:00	163	235	2	0
12678	16134	5	2020	2021-05-23 19:00:00	175	181	0	1
12679	16135	5	2020	2021-05-23 19:00:00	225	179	1	1

12680 rows × 51 columns

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
5270	6018	5	2014	2015-01-18 20:00:00	164	169	2	1
7592	9486	1	2018	2019-03-02 15:00:00	73	88	0	1
11751	15207	3	2020	2020-11-21 14:30:00	262	119	1	2

3 rows × 51 columns

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
5270	6018	5	2014	2015-01-18 20:00:00	164	169	2	1
7592	9486	1	2018	2019-03-02 15:00:00	73	88	0	1
11751	15207	3	2020	2020-11-21 14:30:00	262	119	1	2

3 rows × 51 columns

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
1447	1528	4	2015	2015-11-29 23:30:00	138	146	1	0
2267	2348	5	2016	2016-10-23 22:45:00	161	164	0	0
4644	5392	3	2014	2014-10-18 14:30:00	117	123	6	0
5270	6018	5	2014	2015-01-18 20:00:00	164	169	2	1
5771	7413	1	2017	2018-03-10 15:00:00	219	84	0	0
6179	7821	2	2017	2018-04-17 18:45:00	106	116	4	0
11644	15100	4	2020	2021-04-18 14:15:00	143	156	5	0
12642	16098	5	2020	2021-05-01 19:00:00	160	170	2	0

8 rows × 51 columns

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
1447	1528	4	2015	2015-11-29 23:30:00	138	146	1	0
2267	2348	5	2016	2016-10-23 22:45:00	161	164	0	0
4644	5392	3	2014	2014-10-18 14:30:00	117	123	6	0
5270	6018	5	2014	2015-01-18 20:00:00	164	169	2	1
5771	7413	1	2017	2018-03-10 15:00:00	219	84	0	0
6179	7821	2	2017	2018-04-17 18:45:00	106	116	4	0
11644	15100	4	2020	2021-04-18 14:15:00	143	156	5	0
12642	16098	5	2020	2021-05-01 19:00:00	160	170	2	0

8 rows × 51 columns

```
(12680, 40)
```